Essays on Applied Historical Labour Economics

Rohan Alexander

A thesis submitted for the degree of

Doctor of Philosophy of

The Australian National University

August 2019

Declaration

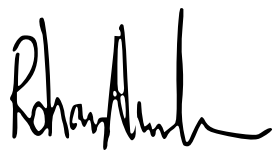
This thesis is an account of research undertaken between February 2015 and February 2019 at the Research School of Economics, The Australian National University, Canberra, Australia.

Chapter 2 was written jointly with Zach Ward and was published as 'Age at Arrival and Assimilation during the Age of Mass Migration' in *The Journal of Economic History*, September 2018. I took responsibility for data gathering, cleaning, and organising; Zach took responsibility for data matching; we jointly worked on the analysis and write-up.

Chapter 3 was written jointly with Tim Hatton. An earlier version of this paper was circulated as part of the Joint APEBH 2019 and All-UC Group in Economic History Conference, February 8 and 9, 2019. I took responsibility for data gathering, cleaning, and organising; we shared responsibility for the model specification and analysis, as well as the write-up.

Chapter 4 is in working paper format, with an intention to submit to a journal after addressing examiner comments.

Except where acknowledged in the customary manner, the material presented in this thesis is, to the best of my knowledge, original and has not been submitted in whole or part for a degree in any university.



Rohan Alexander

12 August 2019

Acknowledgements

Many people generously shared their expertise and time to help and guide my research. In the first instance, thank you to my primary supervisor John Tang who, depending on what was needed, advised, guided, nudged, chided, praised, or encouraged, but was without fail incredibly supportive and very generous with his time. Thank you for your mentorship and friendship John.

My panel was a wonderful source of academic guidance and friendship. Thank you to my chair, Martine Mariotti, who provided invaluable advice both specific to my research papers and more broadly about research practice. Thank you to Zach Ward who involved me in his research program and showed me the process that research goes through from conception to publication. And thank you to Tim Hatton who taught me, by example, how to conduct research in a professional way.

Thank you to the three anonymous examiners for the detailed comments that they provided, which substantially improved this thesis.

Thank you to the broader RSE and ANU community who have been uniformly collegial, welcoming and generous with their time. Thank you especially to Peter Gibbard and Sebastian Wende for their advice and friendship. Thank you to my cohort – Azadeh Abbasi-Shavazi, Ben O'Neill, Minh Ngoc Nguyen, Minhee Chae, and Nabeeh Zakariyya – and the other PhD students, especially Akshay Shanker, James Taylor and Sanghyeok Lee, from whom I learnt so much. Thank you to the many academic staff who were quick to help especially Bob Gregory, Bruce Chapman, Chung Tran, Dana Hanna, Jill Sheppard, Maria Racionero, Tim Kam, Timo Henckel, and Simon Grant. And finally, thank you to the professional staff who went out of their way to help on many occasions,

especially Finola Wijnberg, Jenny Nguyen, Karissa Pereira, Nicole Millar and Susanna Pietrzak.

Much social science research depends on resources put together by others. I am grateful for the extensive resources and datasets that have been made freely available and thank those individuals, governments, and organisations who digitise, clean, administer and maintain these resources.

Thank you to my parents, as well as Mark Alexander and Lauren Falconer, for their support and encouragement. Finally, the partners of PhD students often help create the circumstances in which research can occur, and I certainly thank Monica Alexander for that. But I've been especially lucky to have a partner who can directly advise and guide my research. Monica has variously acted as a fifth advisor, a statistics teacher, an R tutor, an ideas generator, a proof-reader, a co-author, and much more; thank you from the bottom of my heart.

Abstract

The three papers in this thesis reflect original microdata collection and linking that improve how research can be done with historical labour data. In the first paper Zach Ward and I estimate the effect of age at arrival for immigrant outcomes with a new dataset of Ellis Island arrivals linked to the 1940 U.S. Census. Using within-family variation, we find that arriving at an older age, or having more childhood exposure to the European environment, led to a more negative wage gap relative to the native born. Infant arrivals had a positive wage gap relative to natives, in contrast to a negative gap for teenage arrivals. Therefore, a key determinant of immigrant outcomes during the Age of Mass Migration was the country of residence during critical periods of childhood development.

In the second paper Tim Hatton and I examine the votes that led to six British colonies federating to become the Commonwealth of Australia in 1901. We analyse support for Federation using a new dataset of district-level voting records that we associate with a new dataset of district-level census characteristics. We find little support for the view that sectoral interests were important. On the other hand, we find greater support for Federation in districts with a greater share of migrants from outside the colony, among those further from the seats of colonial government, and with a greater share of females. Therefore, support for Federation seems to have been associated more with migration, distance, and possibly female suffrage, than with trade.

In the final, and sole-authored, paper I find that surname analysis suggests a low level of social mobility in Tasmania over the nineteenth and twentieth centuries. Specifically, newly constructed microdata records suggest that the levels of various markers of status between generations are persistent. Surnames are drawn from birth records, while status

is signalled by membership of certain groups, such as being a parliamentarian or attending a certain school in the nineteenth century, and being awarded an Order of Australia or in the legal profession in the late twentieth century. Therefore, social status in Tasmania appears to be correlated over multiple generations.

Contents

Chapter 1

Introduction

The papers in this thesis are about various topics in applied historical labour economics and each of them reflect original microdata collection and linking that improve how research can be done with historical labour data. Each paper is independent, but they share common themes and methods. In the first paper a new dataset of Ellis Island arrivals is constructed and then linked to the 1940 U.S. Census. In the second paper, a new dataset is constructed of district-level support for Australia's Federation in 1901, and this is linked to a new dataset that is constructed of district-level census variables. Finally, in the third paper, various historical and modern individual-level datasets are created and then linked using surnames. In this introduction we summarise each paper, and then discuss their shared themes and methods.

In Chapter 2, Zach Ward and I examine the factors influencing the economic outcomes of immigrants to the United States during the Age of Mass Migration (roughly 1850 through to 1914). We first gather and clean a dataset of Ellis Island arrivals between 1892 and 1924 to identify brothers who arrived at the same time. We then link this dataset to the 1940 U.S. Census to examine the outcomes of more than 50,000 brothers. We focus on brothers because they tend to be subject to the same household-invariant unobservable characteristics and are less likely to change their surname at marriage allowing higher matching rates.

We examine the ages of immigrating brothers when they arrived at Ellis Island, and then we exploit the difference in the ages of these brothers to find that arriving at an

older age led to a more negative wage gap relative to those born in the United States. The gap for a 16-year-old arrival is equivalent to two fewer years of education. We then consider the channels for the effect, and find that while education itself was important other aspects such as imperfect transfer of pre-migration human capital, and reduced social assimilation also matter. The re-emergence of nationalism in the past few years reinforces the relevance of this paper. Our work suggests that the integration of migrants into the receiving country at earlier ages will result in better outcomes regardless of nativist opinions as to their 'quality'.

In Chapter 3, Tim Hatton and I examine why the Australian colonies voted to federate. In 1898 and again in 1899/1900, the six separate colonies conducted referendums on whether to join the proposed federation. We create datasets of votes and census variables, and then link these two datasets by changing the geography of the voting data to match the census records using maps that we digitised.

We analyse the votes at a district level and associate support with a range of district-level census characteristics. We find little support for the view that sectoral interests were important, as would be suggested by the theory of customs unions. Instead we find that the share of people born in a different Australian colony or overseas is a key explanatory variable. Another is the share of females in a district. As those shares increase, support for federating also increases.

In Chapter 4, I examine how social status is passed through generations in Tasmania, one of Australia's oldest states. I first gather and compile a dataset of Tasmanian surnames in the 1800s through to today and some associated memberships, elite school attendance, and occupations. I then follow Clark (2014) and compare the prevalence of certain surnames in certain 'high-status' professions in the nineteenth century, with the prevalence of those surnames today.

I find that 'high-status' prevalence has a high level of persistence over time. The extent to which status is found to be an inheritance in this paper motivates the analysis of modern data such as tax records that could inform appropriate types of policies, and could adjust for some of the weaknesses of the approach. Even if social mobility has

dramatically improved in recent years, there is still a strong cumulative effect across generations.

The main theme of this thesis is the use and linking of larger historical datasets. For instance, in Chapter 2, linking Ellis Island records from between 1892 and 1924 with 1940 U.S. Census records allow us to examine individuals at two points in time. By linking two historical datasets we are able to broaden the types of questions that can be examined, compared with relying on a single time series. Similarly, in Chapter 4, where I examine social mobility in Tasmania, there is no single time series that would allow this to be examined and it is only by linking datasets that this can be explored.

In Chapters 2 and 4 the most important variable for the linking an individual's name. However Chapter 3 uses voting data, for which datasets that include a person's name are rarely available, and historical Australian census records, for which individual level responses are not available. Instead we link the two datasets using geographies. After digitising maps of census areas we are able to identify the census area of each voting booth and then construct voting outcomes for census areas.

This thesis illustrates the benefit of constructing linked datasets. This allows the examination of phenomena that take decades or centuries to evolve without needing unbroken time series. As digitisation technologies continue to improve the methods used in this thesis will able to used in an increasingly large number of areas.

Chapter 2

Age at Arrival and Assimilation During the Age of Mass Migration

2.1 Introduction

It is increasingly apparent that where one was born and the quality of one's childhood environment are key determinants of life-long outcomes.[1] By definition, immigrants are born in a different environment than natives; therefore, immigrants are exposed to a different educational, cultural, and health setting during critical periods of development. How much of the economic gaps between immigrants and natives during the Age of Mass Migration can be attributed to growing up in different environments? There are many other factors that may explain the gaps between immigrants and natives besides where one grew up, such as the direction of selection into immigration, the degree of discrimination from natives, or the extent of sorting into different enclaves (Biavaschi et al. (2017); Borjas (1987); Cutler et al. (2008)).

To estimate the importance of growing up abroad we exploit variation in the length of childhood exposure to source country conditions, as measured by the migrant's age at arrival. By comparing the adult outcomes of older child arrivals to younger child arrivals,

[1] The effect of childhood environment on economic outcomes is a long-standing question in the economics literature. For recent literature reviews, see Almond et al. (2017) and Cunha et al. (2006) on the importance of environment during early stages of childhood. Also see the work on the importance of childhood environment past age eight (Chetty and Hendren (2017a); Chetty and Hendren (2017b); Chetty et al. (2016)).

we can uncover the extent to which outcomes in the United States depended on where one spent his infancy or adolescence (Chetty and Hendren (2017a); Chetty and Hendren (2017b)). This method also allows us to identify critical ages for when a move improved migrant outcomes the most; prior research on child arrivals misses variation within the group by treating all children as a single category (Hatton (1997); Minns (2000)).

We take advantage of the complete digitisation of immigration records to construct a sample of brothers arriving at Ellis Island between 1892 and 1924, which we then link forward to the full-count 1940 Census. With a linked dataset of more than 50,000 brothers, we then estimate the effect of age at arrival by comparing brothers who immigrated at different ages. This strategy controls for household-invariant unobservable characteristics such as parental income and education that may be correlated with both age at arrival and migrant outcomes (Böhlmark (2008); Van den Berg et al. (2014); Clarke (2016)).

We find that an older age at arrival, and thus longer exposure to the childhood environment in Europe, had a large and negative effect on the native-immigrant gap in outcomes such as wage income and occupational status. For 16-year-old arrivals, the native-immigrant wage gap was 17 log points more negative compared with the gap for those who arrived at age one—an effect that is equal in size to two fewer years of education. The size of this effect is larger than the overall wage gap between teenage arrivals and white natives; therefore, we show that infant arrivals had a positive wage gap relative to natives, in contrast to a negative gap for teenage arrivals.

After establishing that arriving at an older age had a large negative effect on the native-immigrant gap in economic outcomes, we explore potential channels for this effect. One mechanism is through educational attainment: 16-year-old arrivals acquired one less year of schooling than infant arrivals. However, a one-year difference in education does not explain the entire income effect, suggesting that other mechanisms besides educational attainment were important. We show that older arrivals were also penalised because potential foreign labour market experience was not rewarded in the United States, implying that pre-migration human capital did not transfer perfectly across borders. Older arrivals were also less socially assimilated, as measured by their rate of marriage to a native-born

spouse, which may have penalised them in the labour market (Abramitzky et al. (2016); Biavaschi et al. (2017)). While we cannot pinpoint which channel was most important, we consistently show that longer exposure to the European environment during critical periods of development was strongly correlated with a variety of migrant outcomes during the Age of Mass Migration.

Our study contributes to the growing literature on immigrant assimilation during the Age of Mass Migration using newly digitised records (Abramitzky et al. (2014); Abramitzky et al. (2016); Biavaschi et al. (2017); Ward (2018)). The current understanding in the literature is that the average immigrant's position in the occupational distribution was fixed and did not change relative to natives throughout the life cycle; note that this does not imply income convergence did not occur, but incomes are unobserved prior to 1940. We show that a male immigrant's position in the occupational distribution depended strongly on his age at arrival. These results suggest that while human capital acquired during adulthood, such as English fluency post arrival, had a smaller impact on the native-immigrant gap in occupations, human capital acquired during childhood had a larger impact (Ward, 2018). Our results also add to the growing literature on age-at-arrival effects by showing that they were large and important during the Age of Mass Migration, a time period when the economic gap between source countries and the United States was smaller than the economic gap between source countries and the United States today (Abramitzky and Boustan (2017); Böhlmark (2008); Van den Berg et al. (2014)).

Our study also complements the literature on the intergenerational assimilation of immigrants (Abramitzky et al. (2014); Borjas (1994); Card et al. (2000)). Child immigrants are sometimes called the "1.5" generation since they bridge the gap between adult arrivals in the first generation and native-born individuals in the second generation. Our results suggest that the second generation should have improved on the first generation's relative position with natives since the second generation spent their entire childhood within the United States. Yet the intergenerational assimilation literature also documents that convergence of occupational status for descendants from different source

countries was not complete for the children and grandchildren of European migrants, even though these generations were raised in the same country. A lack of convergence across generations from different sources is consistent with there being large differences in the quality of childhood environment across areas within the United States where immigrants from different sources settled, a potential topic for future research.

2.2 Historical Setting And Related Literature

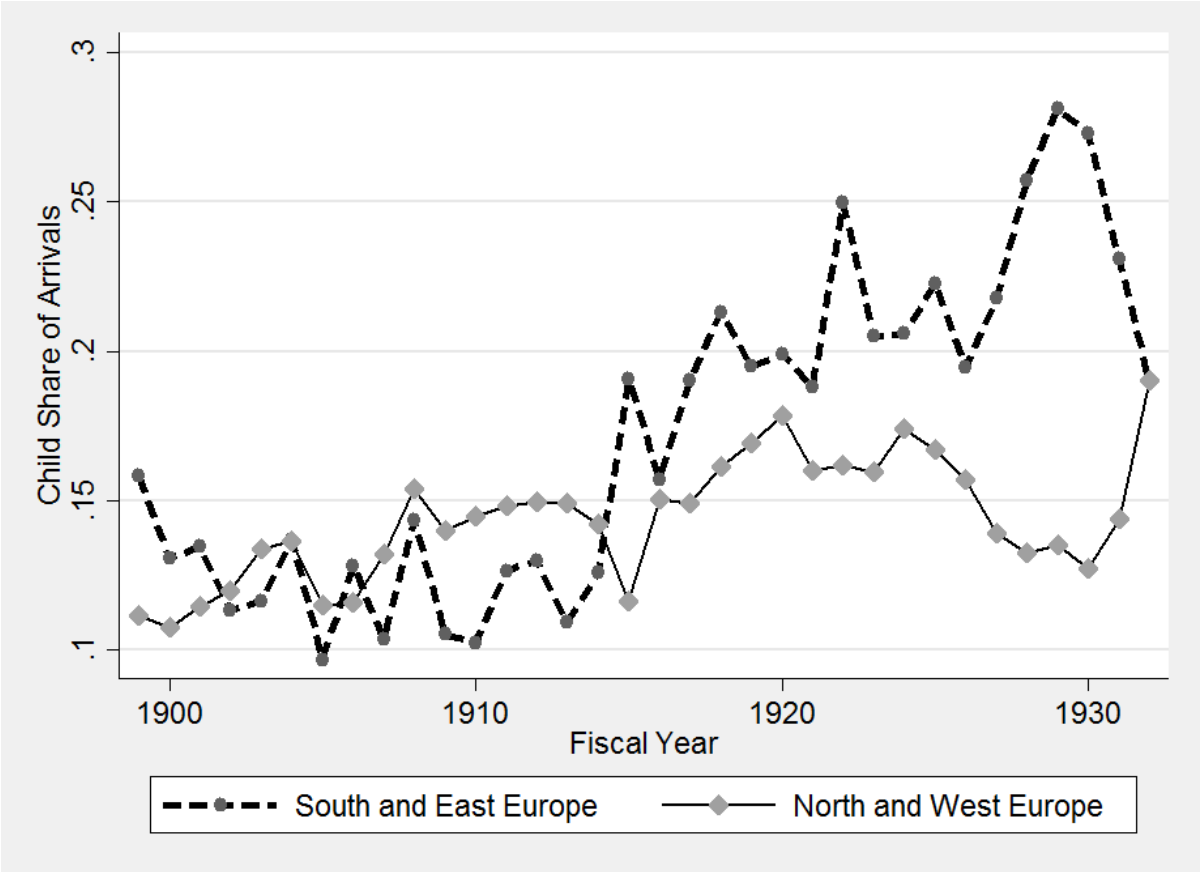
The Age of Mass Migration (1850–1913) is often split into two sub-eras based on the geographical shift of flows from Northern and Western Europe ("Old" sources) to Southern and Eastern Europe ("New" sources) in the late 1880s. At around the same time there was also a shift in family composition from intact households (including many children) to unattached males, lowering the fraction of child arrivals (Baines (1995); Hatton and Williamson (1998)). Illustrating this shift in the late nineteenth century, Greenwood (2007) reports that the percentage of those under 14 in the inflow from major European sources dropped from a high of 25 per cent in 1884 to a low of 11 per cent in 1895. This shift away from family and child migration is associated with younger males taking advantage of the decreasing costs of travel due to the diffusion of steam technology and migration networks (Cohn (2009); Gould (1980)).

Our data cover migrants who entered through Ellis Island between 1892 and 1924, a period when child arrivals were slowly making up a larger share of arrivals (see Figure 2.1).[2] Children were still a small, but increasingly important, part of the inflow: overall, the fraction of child arrivals increased from 10 to 14 per cent before WWI to slightly above 15 per cent in the following decade. Some of this increase is due to several shocks

[2]The data in the series are from the Annual Reports of the Commissioner General of Immigration (1899–1932). One caveat to Figure 2.1 is that both the definition of an immigrant and a child arrival changed during the early twentieth century (Hutchinson, 1958). Prior to 1903, any entrant, excluding the cabin class, was counted as an immigrant. For the following two years (1904 and 1905), the definition changed to include the cabin class. From 1906 onward immigrants were those who intended to stay for more than one year and had been outside of the United States for more than one year. Besides the definition of immigrant, the definition of a child arrival also changed from those under the age of 14 prior to the 1917 literacy test to those under the age of 16 afterwards. A final caveat is that the Annual Reports may have underestimated the number of arrivals due to careless compiling of ship manifests by the Bureau of Immigration (Bandiera et al., 2013). However, since undercounting is mostly due to entire ships missing from the totals, it is unclear how it would bias the fraction of children in the arrival flow.

to the immigration system, such as the cut-off of flows during WWI, the Literacy Act of 1917, and the immigration quotas laws of 1921 and 1924. While U.S. policy significantly restricted the overall flow, child arrivals were favoured under these policies since those under 16 were not subject to the literacy test, and children joining a naturalised family member were given preference under the quota system.[3] Consistent with policies favouring children more than single adults, the countries that were more restricted under the quotas and literacy test (in Southern and Eastern Europe) had a relative increase of children in their flow.

Figure 2.1: Child Share Of Immigrant Inflows To The United States, 1899–1932



Notes: Fiscal years are between 1 July and 30 June. Child arrivals are those under the age of 14 between 1899 and 1916 and under the age of 16 between 1917 and 1932. See Footnote 3 for definition of arrival.
Sources: Annual Reports of the Commissioner General of Immigration, 1899–1932.

While child arrivals were less than 20 per cent of arrivals in the early twentieth century, they were about 30 per cent of the migrant stock, partially because they were more likely

[3]Both the 1921 and 1924 immigration quotas allowed child immigrants to join naturalised family members even if the quota for the country was full.

to remain rather than return home.[4] This can be directly seen in return flow records where children were underrepresented on out-going ships relative to the migrant stock; moreover, arrival records show that families with children were more likely to plan to stay in the United States permanently than single arrivals (Ward, 2017). Yet not all young migrants arrived with family members; this is indirectly seen in the distribution of age at arrival in the migrant stock in Figure 2.2.[5] While there were about the same proportion of arrivals at age one as for age 12, there were much more arrivals aged 13 and above, perhaps because teenagers were more likely to migrate by themselves. If older arrivals came individually while younger arrivals came as part of a family, then older teenagers could be selected from a different part of the source country human capital distribution; therefore, it will be important to estimate the effect of age at arrival with an empirical strategy that accounts for changing unobservables across the arrival age distribution.

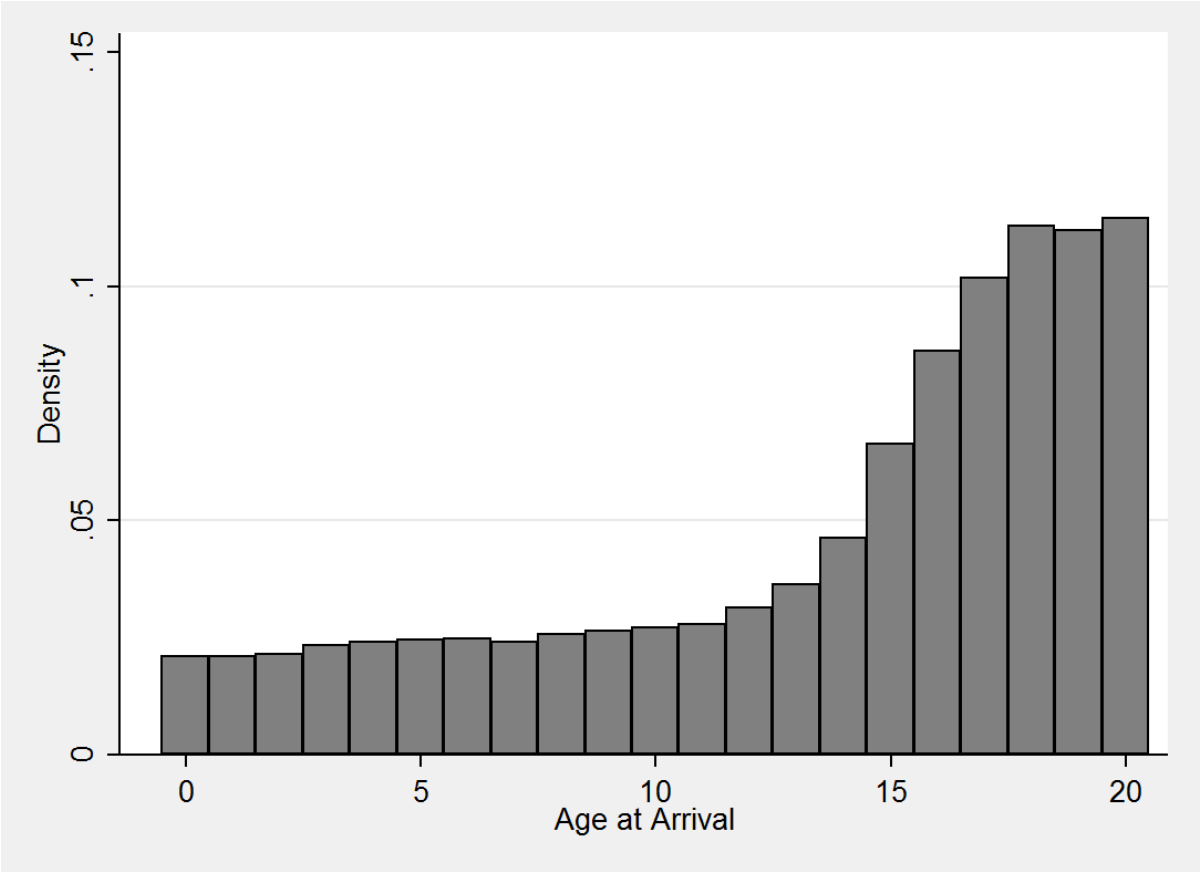
Early twentieth century officials recognised the importance of age at arrival for successful assimilation; of special concern was whether older arrivals were falling behind in school. The 1910 Dillingham Commission Reports on The Children of Immigrants in Schools showed that 43 per cent of children who arrived under age six were behind their grade level, compared with 92 per cent of those who arrived at ten years or older. The authors argued that "the child who comes to this country before he reaches school age often has an opportunity to adjust himself to his new surroundings and in some cases learn the language through contact with other children before entering school" (US Immigration Commission (1910), p. 51). In response to this trend of child arrivals being poorly educated, states passed compulsory schooling laws to educate immigrant children who arrived from countries without a compulsory educational system (Bandiera et al., 2016).

Despite early twentieth century officials' interest in age at arrival, the Congressional

[4] This 30 per cent number is based on 1899 to 1930 arrivals in the 1900–1930 IPUMS samples (Ruggles et al., 2017).

[5] This figure is created using the 1900-1930 IPUMS samples (Ruggles et al., 2017) and keeping those who arrived between 1899 and 1930 to match Figure 2.1. A random sample of ships to Ellis Island from (Ward, 2017) confirms that older arrivals tended to travel alone, where about 20.5 per cent of 14-year-olds, 23.6 per cent of 15 year-olds, and 52.9 per cent of 16 year olds entered the United States without a family member (defined by same surname) on the ship.

Figure 2.2: Distribution of immigrant age-at-arrival in the 1900-1930 U.S. censuses



Notes: The sample is limited to those who arrived between 1899 and 1930 to match with Figure 2.1. Distribution is estimated after applying the person weight available from IPUMS.
Sources: 1 per cent samples of the 1900–1920 Censuses, 5 per cent sample of the 1930 Census (Ruggles et al., 2017).

Report is one of the only studies that separates historical migrant outcomes by arrival age.[6] Others that account for age at arrival often group all child arrivals into a single category. Both Minns (2000) and Hatton (1997) show that those who arrived under the age of 16 had higher income levels and better-paid occupations than adult arrivals, consistent with a negative effect of age at arrival and longer exposure to the European environment. On the other hand, Abramitzky et al. (2014) show that assimilation rates were similar whether one keeps or drops those who arrived under the age of ten, but since they do not isolate the sample to only child arrivals, the difference in assimilation for child

[6] Ward (2018) estimates the effects of age at arrival on English proficiency using the 1900 to 1930 U.S. cross sections, and indicator variables for each arrival age. Ward is primarily interested in using the estimates to verify the quality of the English proficiency variable rather than to directly analyse the effect of age of arrival on occupational outcomes.

arrivals is unclear. We improve on this limited literature by estimating the effect of age at arrival across all ages, rather than grouping all child arrivals together. We also use an empirical strategy that controls for household-invariant unobservables that are correlated with age at arrival and with economic outcomes, which is important in today's studies on age at arrival (Clarke, 2016).

In contrast to the scarcity of historical studies, several modern-day studies estimate the effect of age at arrival on adult outcomes with high-quality data.[7] The most credible method to identify the age-at-arrival profile uses sibling fixed effects. This requires a large amount of data and therefore has been primarily studied using Swedish and Norwegian administrative records (Böhlmark (2008); Van den Berg et al. (2014)). Outside of Northern Europe, there are few studies that identify the effect of age at arrival with siblings. (Chetty and Hendren, 2017a) and (Chetty and Hendren, 2017b) use U.S. tax records to show that variation in age at migration across counties has a large effect on adult outcomes, implying that childhood environment varies widely across counties in the United States. We follow this sibling fixed effects approach to estimate the importance of childhood environment for immigrants from the past.

2.3 Linking Ellis Island Records To The 1940 Census

The main dataset used for estimation comes from linking two large data sources: Ellis Island records from 1892 and 1924 and the preliminary full-count 1940 Census available at IPUMS (Ruggles et al., 2017). The Ellis Island records have been digitised and are searchable online; note that this source is the same one used by Bandiera et al. (2013) and Spitzer and Zimran (2017).[8] While the clear advantage of the Ellis Island records is that they include millions of observations, there are a few disadvantages. One is that not every variable in the arrival records is digitised, such as occupation, relationship status, or

[7] Friedberg (1992) is the seminal study of age at arrival on adult outcomes. Several outcomes besides income have been explored, including human capital outcomes such as language acquisition and educational attainment (Bleakley and Chin (2004); Böhlmark (2008); Schoellman (2016)), social outcomes such as intermarriage or living in an ethnic enclave (Åslund et al. (2015); Bleakley and Chin (2010)), and health outcomes such as height (Van den Berg et al., 2014).

[8] Many arrival records prior to 1897 were lost in a fire, so coverage prior to 1897 is not complete (Spitzer and Zimran, 2017).

height. Moreover, the Ellis Island records include both immigrants and non-immigrants; non-immigrants are other entrants such as business travellers, tourists, or those traveling through to another country. However, we are only interested in those who we can locate in the 1940 Census, and thus those who have stayed permanently (and survived) until 1940.

Our population of interest in the Ellis Island records is brothers who are single and arrived between the ages of zero and 20. For this population (who are primarily European), we collect first name, last name, age, date of arrival, place of last residence, and ethnicity. We identify brothers as immigrants who are listed next to each other with the same last name and are less than ten years apart in age, although we do not have their relationship listed in the data.[9] The key variable of interest from these records is age, which we wish to attach to their adult observation in the 1940 Census. For a discussion of the assumptions we made in cleaning the data, please see Appendix A.2. After cleaning, we have 397,003 brothers who can be linked to the 1940 Census.

We link these brothers to the 1940 U.S. Census using a match on first name, last name, country of birth, and year of birth in a 3-year range.[10] We find potential matches based on having an exact NYSIIS match on first and last name; however, we choose the best match based on the smallest sum of the absolute difference in year of birth, Jaro-Winkler distance in first name and Jaro-Winkler distance in last name.[11] (Massey, 2017) shows that this method of ranking matches is reasonable for improved match rates and reduced false positives. For more detail on the linking process, see Appendix A.3.

It is possible that our linking methodology incorrectly links some people, which would induce measurement error and bias our sibling fixed effects estimates toward OLS esti-

[9] The two people must have also arrived at the same time.

[10] We access the 1940 Census on the National Bureau of Economic Research (NBER) server due to restrictions on observing the first name and last name in the public-use dataset. Some of the data files created in this study are available online (Alexander and Ward, 2018). However, the linked records have not been made available since the 1940 Census is restricted access.

[11] NYSIIS, or the New York State Identification and Intelligence System, is a phonetic algorithm to standardise similar sounding names. The Jaro-Winkler algorithm measures the distance between strings based on the number of matching characters. Using the actual first and last name strings to gauge the quality of match is recommended by Bailey et al. (2017), rather than treating all matches with the same NYSIIS code as of equal quality. We show that results are robust to using a method related to Feigenbaum (2016) in Appendix A.5.

mates (Bailey et al., 2017). However, as we will show in robustness checks, our results do not change when limiting our sample to higher-quality links in terms of closer matches in first and last name strings and year of birth. The results are also robust in an alternative sample where links are chosen based on a predicted match score calculated from a hand-linked sample of immigrants, a method that is related to the linking strategy described by Feigenbaum (2016).[12] Overall, we are confident that our results are not driven by link quality.

We take one extra step when linking the datasets because we start with arrival records unlike others who link from census to census. We are concerned that some immigrants may have changed their first name to be more "American" after arrival; for example, from Giuseppe to Joseph or Pietro to Peter. Biavaschi et al. (2017) show that name changes occur for 32 per cent of their sample of naturalisation records in New York and that name changes were more common for Southern and Eastern Europeans. To account for this possibility, we Americanise the first names in our dataset of arrival records and the first names in the census records. This will allow us to match Giovanni at arrival to John in the census, but also to match Giovanni (Americanised to John) at arrival to Giovanni (Americanised to John) in the census in case Americanisation did not occur. We do this with a list of more than 28,000 variants of first names based on information at `behindthename.com`.[13] The Americanisation process improves our linking rates by about 35 per cent, but, as we will show in a later robustness check, our results do not substantially change if we do not Americanise first names.

The starting sample of 397,137 brothers is successfully linked for 103,005 individuals in the 1940 Census using our main linking approach, or 25.9 per cent of arrivals. Since the main empirical strategy exploits variation within brothers, we drop individuals where one brother was linked and another was not. This restriction gives us a final sample of

[12] We use the hand-linked samples from Ward (2018) to predict the best link among the set of potential links. While this method is related to Feigenbaum (2016), it is not exactly the same since our "training sample" is from immigrants linked between 1920–1930 U.S. Censuses rather than Ellis Island records to the 1940 Census.

[13] For some names, there are multiple American-sounding variants. We choose the variant that is most popular for years of birth prior to 1930, data which is available from the Social Security Administration at `http://www.ssa.gov/oact/babynames/names.zip`.

53,129, or 13.4 per cent of our original set of brothers.

Table 2.1 shows the linking rates by country of birth and demonstrates a common pattern in the literature where we are less likely to link Southern and Eastern Europeans relative to Northern and Western Europeans (Abramitzky et al. (2014); Ward (2018)). Clearly, our linked sample is not a random sample of foreign-born brothers. We are not able to test for the representativeness of the sample on occupation or literacy compared to all Ellis Island arrivals since these variables are not digitised. Yet we would rather test for representativeness according to the 1940 Census since the Ellis Island records include many non-immigrants and thus any difference between our linked sample and those in the Ellis Island data would reflect both selection into permanent migration and selection into the linked sample. However, we also cannot test for representativeness according to the 1940 Census because it does not separate immigrants by cohort or age of arrival, once again making it unclear whether any differences are due to biases from the linking process or because the linked data has younger arrivals.[14]

Most linked samples that use a similar linking methodology are found to be slightly higher skilled than the underlying population and only show a strong bias in country of birth (Abramitzky et al., 2014). New source countries are less likely to be linked to the 1940 Census than Old sources because of common names, return migration, misspelled names, or names that are not captured in our list of Americanised names. To account for this bias, we reweight the sample to reflect the migrant stock by country of birth in 1940, although this reweighting does not drive our results.[15]

Table 2.2 shows the descriptive statistics of the linked sample of brothers and the 1 per cent sample of white natives in the 1940 census, illustrating the gaps in economic and social outcomes between immigrants and natives. We compare our immigrants to white natives on several outcomes, including years of completed education, occupation,

[14]We can compare our sample to the 1940 migrant stock, which we do in Appendix A.2. Our linked sample is higher skilled, more highly educated, and less likely to be from a new source country.

[15]Reweighting to match the 1940 stock is done with males who were born between 1872 and 1924 to reflect our sample of zero to 20-year-old brothers who arrived between 1892 and 1924. We alternatively reweighted to match the 1930 Census distribution of country of birth, a census which includes year of arrival and thus we can reweight to match those who arrived between 1892 and 1924. Either weighting to match the 1930 or 1940 migration distribution yields the same results.

Table 2.1: Birthplace composition in Ellis Island data linked to the 1940 census

Country of Birth	Brothers at Arrival	Linked to Census	2+ Brothers Linked	2+ Brothers Link Rate
Old source countries:	96,790	34,877	21,421	22.1
Denmark	3,876	1,558	961	24.8
Finland	3,807	677	269	7.1
Norway	7,468	2,363	1,306	17.5
Sweden	8,621	3,226	1,816	21.1
England	23,695	9,601	6,091	25.7
Scotland	9,739	5,109	3,512	36.1
Ireland	8,154	4,433	3,130	38.4
Belgium	3,694	653	272	7.4
France	5,668	843	332	5.9
Netherlands	11,752	2,680	1,369	11.6
Switzerland	2,848	898	528	18.5
Germany	7,468	2,836	1,835	24.6
New source countries:	275,205	65,952	30,799	11.2
Greece	9,411	1,139	298	3.2
Italy	150,476	49,183	24,224	16.1
Portugal	2,376	740	445	18.7
Spain	2,812	590	291	10.3
Austria	14,789	2,251	895	6.1
Czechoslovakia	9,835	2,309	1,157	11.8
Hungary	8,736	1,204	450	5.2
Poland	16,717	2,204	794	4.7
Romania	6,757	665	197	2.9
Yugoslavia	3,309	572	205	6.2
Lithuania	2,759	122	22	0.8
Russia	47,228	4,973	1,821	3.9
Other (Asia, Canada, Mexico)	25,142	2,176	909	3.6
Total	397,137	103,005	53,129	13.4

Notes: The empirical strategy uses sibling fixed effects, so we only keep sets of brothers where at least two are successfully linked (2+ Brother Linked column). For context, the resident population of the United States at the 1940 US Census 132 million people.
Sources: Linked sample of brothers from Ellis Island records to the 1940 Census.

and wage income. Note that whenever we use wage income, in this table or in later regressions, we exclude self-employed workers since business and farm income are not included in the 1940 Census.

The migrants in our sample have been in the United States for an average of about 31 years. Therefore, those who arrived between age zero and five are on average 34 years old in 1940, while those who arrived between ages 16 and 20 are on average 49 years old in 1940. Considering these differences in age, it will be important to adjust for age when

Table 2.2: Descriptive statistics of linked sample of brothers

		Age at arrival			
	Native-born	0–5	6–10	11–15	16–20
Age	35.76	34.80	39.50	43.52	48.58
	(14.58)	(7.815)	(7.752)	(7.815)	(7.161)
Years in the United States		31.91	31.47	30.76	30.85
		(7.654)	(7.652)	(7.755)	(6.983)
Southern and Eastern		0.544	0.601	0.625	0.623
European (New source)		(0.498)	(0.490)	(0.484)	(0.485)
Years of U.S. education	9.286	8.169	5.615	1.565	0.165
	(3.280)	(3.392)	(3.616)	(2.378)	(0.789)
Years of foreign education		0	1.861	5.157	5.987
		(0)	(1.389)	(2.327)	(3.369)
Potential U.S. labour	20.49	20.60	25.83	29.17	30.66
market experience	(15.50)	(9.515)	(9.363)	(8.405)	(7.078)
Potential foreign labour		0	0.163	1.605	5.741
market experience		(0)	(0.652)	(2.418)	(3.570)
		Age-adjusted difference from white native-born			
Log (income), if wage	6.712	0.0945	0.0344	–0.00257	–0.0792
worker	(0.960)	(0.686)	(0.697)	(0.683)	(0.722)
Log (income), if wage	6.905	–0.0428	–0.108	–0.153	–0.225
worker and urban	(0.878)	(0.677)	(0.687)	(0.671)	(0.704)
Self employed	0.226	–0.0587	–0.0585	–0.0631	–0.0870
	(0.418)	(0.359)	(0.389)	(0.410)	(0.418)
White collar	0.291	–0.0437	–0.0472	–0.0786	–0.107
	(0.454)	(0.444)	(0.447)	(0.430)	(0.408)
Farmer	0.129	–0.0832	–0.102	–0.109	–0.123
	(0.336)	(0.180)	(0.182)	(0.210)	(0.231)
Unskilled	0.418	0.0822	0.105	0.137	0.187
	(0.493)	(0.501)	(0.502)	(0.502)	(0.502)
Semi-skilled	0.162	0.0447	0.0444	0.0504	0.0432
	(0.368)	(0.411)	(0.418)	(0.426)	(0.423)
Urban area	0.532	0.255	0.256	0.261	0.255
	(0.499)	(0.395)	(0.396)	(0.399)	(0.413)
Native-born spouse,	0.960	–0.317	–0.391	–0.483	–0.628
if married	(0.196)	(0.477)	(0.494)	(0.498)	(0.468)
Fraction of HH in county	0.838	–0.145	0.148	–0.148	–0.149
which are native born	(0.162)	(0.141)	(0.138)	(0.137)	(0.137)
Observations	372,870	14,979	15,693	11,367	11,090

Notes: Native born are white males restricted to the same birth cohorts are the immigrant sample. Total education and potential labour market experience is split under the assumption that individuals enter school at age six and continuously attend for their full years of schooling (see Footnote 19). The outcomes are age-adjusted residuals after predicting life-cycle variation with the native born (see Equation 2.1 in text). There is missing information for some of these variables. Specifically, 17,152 do not have a positive log income, 1,538 have missing education, 5,226 have missing self-employment, and 3,225 have a blank occupation. HH stands for household heads.
Source: Linked sample of brothers from Ellis Island records to the 1940 Census and a 1 per cent sample of 1940 Census.

examining differences between immigrants and natives, which we show in the bottom half of the table. After adjusting outcomes based on white natives' life-cycle profile, Table 2.2 shows that there is a strong negative gradient to age at arrival for many variables.[16] For example, zero-year-old arrivals earned 9.5 per cent more than natives, while 16–20 year old arrivals earned 7.9 per cent less. Note that children arriving early enough have earnings that are higher than natives of the same age, perhaps implying that childhood environment may explain the entire native-immigrant wage gap for older arrivals. Yet part of the reason why younger arrivals earned more than white natives overall was because they located in urban areas. Table 2.2 shows that when limiting the sample to those only in urban areas, then zero-year-old arrivals earned 4.3 per cent less than natives. Nevertheless, when limiting the sample to urban areas, the same pattern holds where older arrivals had a larger wage gap with natives compared with younger arrivals.

One explanation for the change in income gap across age at arrival is that older arrivals were exposed to source country conditions for a longer period. On the other hand, the change may be due to selection bias such that older arrivals had worse outcomes because they came from lower income or educated families; the 'push' and 'pull' factors may have changed. Instead of estimating the age-at-arrival profile using variation across families, we will estimate the profile using variation within family to control for unobserved family-invariant variables (such as parental education and income) as described in more detail in the next section.

2.4 Empirical Strategy And Identification

The first challenge when estimating the effect of age at arrival on immigrant outcomes is a standard one of collinearity: it is not possible to simultaneously estimate the effect of age at arrival, age, and years in the United States because they are linearly dependent.[17] We follow the standard practice of using natives to identify the life-cycle profile

[16] To adjust for age, we use the standard method in the assimilation literature and run a first regression of the outcome on the full-range of age fixed effects with our sample of white natives, and then calculate the residuals for the sample of immigrants based on predicted values from natives. See Equation 2.1 and the dependent variable of Equation 2.2 in the next section.

[17] That is Age at Arrival = Age – Years in the United States.

(or aging effect), and then estimate whether age at arrival influences deviations from this profile (Borjas, 1985).[18] We take the two-step approach used by Schaafsma and Sweetman (2001): first, we estimate an auxiliary regression to identify the age-earnings profile using only white native-born individuals (superscript nb) from the same birth cohorts as our immigrant sample:

$$ln(\text{income}) = \lambda_a^{nb} + \epsilon_i. \tag{2.1}$$

The income-age profile is modelled using a full-range of age fixed effects. We then estimate whether deviations from the native age-earnings profile are related to the immigrant's age at arrival, while controlling for other factors such as years in the United States:

$$ln(\text{income}) - \hat{\lambda_a^{nb}} = g(\text{Age At Arrival}_i) + h(\text{Years in US}_i) + \upsilon_i. \tag{2.2}$$

Therefore, this equation shows that the estimated effect of age at arrival on outcomes is the effect of age at arrival on the native-immigrant gap.[19] The method essentially estimates the effect of differences in age at arrival on the difference between natives' and immigrants' log income by age.

The second problem with estimating the effect of age at arrival is selection bias: immigrants who arrive at older ages may differ from those who arrive at younger ages in unobservable ways. For example, families with a strong preference for improving their child's education may have immigrated to the United States with children at younger ages prior to school entry. In this case, an estimated age-at-arrival effect may capture family preferences for investment into children and lead to a negative age-at-arrival profile if families with younger children are positively selected relative to families with older children. Indeed, in present-day data, those immigrating with younger children also tend to have higher education levels than those migrating with older children (Clarke, 2016).

To address issues of selection bias when comparing immigrants across families, we compare immigrants within the family (i.e., we compare brothers). The regression therefore

[18] We use the preliminary full-count 1940 Census to estimate the life-cycle profile. We use male white native-born who are aged 15 to 69 to match the immigrant sample. Wage income is top coded at 5,000.

[19] Note that when estimating this equation, the years in the United States function is a mix of assimilation and cohort effects since we only have a cross section (Borjas, 1985).

changes to

$$ln(\text{income}_{ih}) - \hat{\lambda_a^{nb}} = g(\text{Age At Arrival}_{ih}) + \sigma_h + \upsilon_{ih}.$$

where the key addition is the sibling fixed effect σ_h. Therefore, we relate the variation of native-immigrant income gap within siblings to the variation in age at arrival within siblings. Including household fixed effects controls for many household-invariant factors such as parental preferences for education or childhood investment, parental wealth, father and mother's education, culture, family structure, and country of origin. Note that this strategy compares individuals from the same arrival cohort with the same number of years in the United States, so these other variables of interest in the assimilation literature are dropped from the equation.

We use a non-parametric approach and code age at arrival into two-year bins (arrived between zero and one, between two and three, etc.) up until arrival at age 18.[20] This specification allows us to capture a variety of slopes in the profile such as the age-at-arrival profile being flat until ages eight to ten and decreasing afterwards, reflecting language acquisition or other effects of this critical period (Bleakley and Chin (2004); Van den Berg et al. (2014)). Alternatively, the slope could be steepest for arrival ages under five, reflecting the importance of human capital development at very young ages (Almond et al., 2017). Note that we do not control for any post-arrival outcome, such as geography or marital status, because location could be an outcome of age at arrival (Bleakley and Chin, 2010).

For the regression to estimate a causal relationship, the identifying assumption is that age at arrival is not correlated with unobservables that vary within the family and also affect income. Unfortunately, we are unable to include other control variables that may bias our estimates due to the limited information in arrival records. The primary concern is birth order: birth order may affect adult outcomes through general birth order effects, and birth order is also correlated with age at immigration. It is also possible that parents

[20] We code the bins to the floor of the two ages such that ages zero to one are bin zero, two and three are in bin two, up to bin 18, which includes 18, 19, and 20 year olds. We include 20-year-old arrivals in this bin due to a small number of observations. We show in a robustness check that excluding ages 16 and up from our sample does not change the results. Moreover, using one-year bins does not change the qualitative conclusions, but increases the noisiness of estimates.

may time immigration to be optimal for the younger child (e.g., immigrate just prior to school entry) such that younger arrivals would have better outcomes due to unobservable parental investment into younger children. Unfortunately, we only observe birth order according to the arrival records, which ignores older siblings who may have stayed in the source country (Abramitzky et al., 2013). Since birth order is not exactly observed, we do not control for it in our main specification; nevertheless, we show in a robustness check that the results are robust to controlling for birth order.

There are a few threats to the external validity of our estimates, where they may not apply to all child immigrants during the Age of Mass Migration. First, we only identify the age-at-arrival profile with brothers rather than single arrivals; these estimates may differ if there is an extra disruptive effect for older brothers who have to take care of their younger brothers, though the sizes of our effects are likely too large for this to be driving our result. Another possible bias is that we mismeasure the effect of age at arrival for the entire population due to selective return migration (Abramitzky et al. (2014); Ward (2017)). Our sample only consists of brothers who remained in the United States; however, if one brother stayed and another returned home, then they would not be included in our sample. If older arrivals were more likely to return home and older arrivals also earned less income, then we would understate the negative effect of age at arrival. A similar story and bias would apply in the case of selective mortality. Our method for identifying brothers means that we also do not identify brothers who arrive separately. Finally, our estimates do not apply to all arrivals since we only link those who arrived through Ellis Island, which primarily misses entrants from Asia, Canada, or Mexico.[21] Therefore, the reader should keep in mind that our results come from brothers who arrived at Ellis Island between 1892 and 1924 and survived until the 1940 Census.

[21] The 1907 Annual Report of the Commissioner General of Immigration lists that about 80 per cent of immigrant arrivals were to New York. This may have decreased following the immigration quotas when more immigrants entered via land borders.

2.5 The Effect Of Age At Arrival On Economic Outcomes

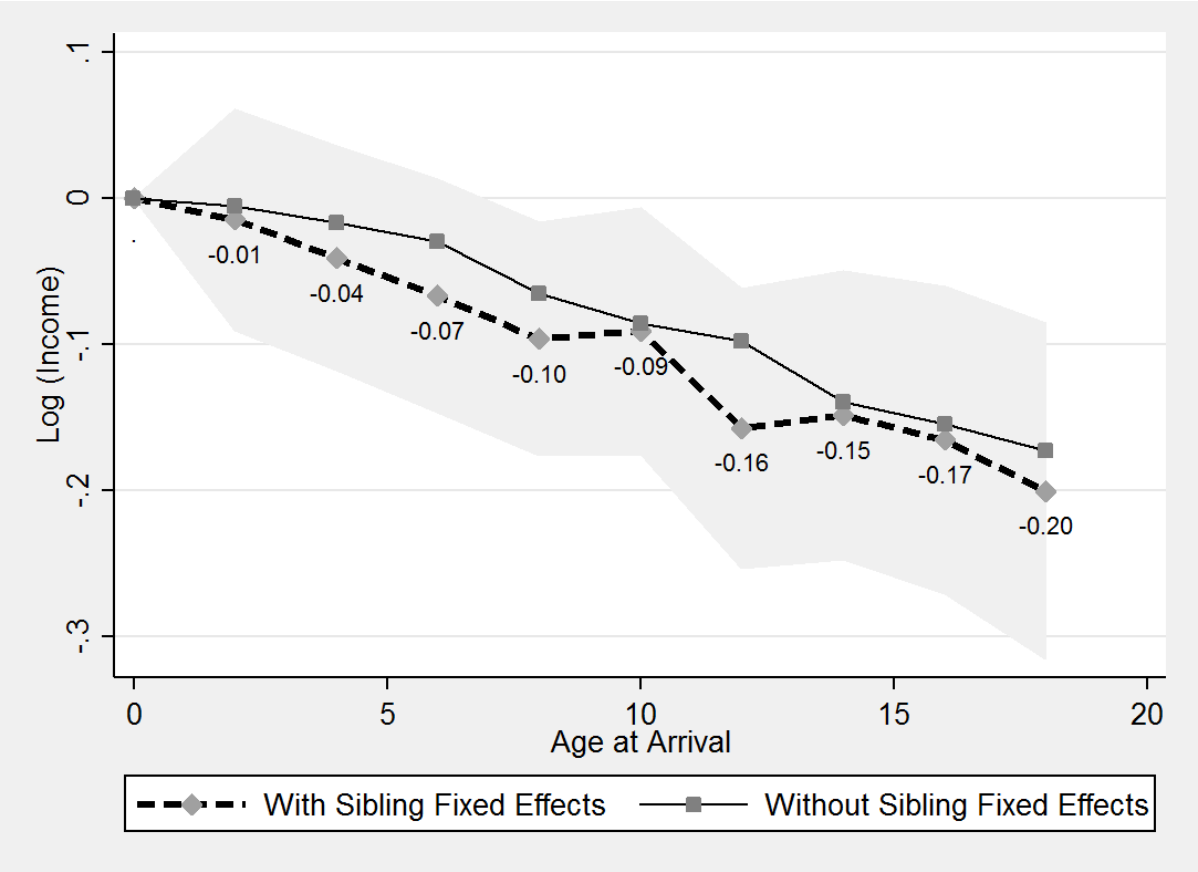
We estimate the effect of age at arrival with the brothers fixed effects specification and plot the coefficients in Figure 2.3.[22] The plotted coefficients estimate the difference in native-immigrant wage gap relative to the native-immigrant wage gap for zero- to one-year-old arrivals. The results show a strong negative slope for age at arrival such that the native-immigrant wage gap was 17 log points (or 15.6 per cent) more negative for a 16-year-old arrival compared with the wage gap for an infant arrival. If one assumes that the return to education was 6.5 to 7.9 per cent in 1940 (Clay et al., 2016), then having a full childhood in the United States was equivalent to receiving more than two additional years of schooling.

This result is consistent with the U.S. childhood environment yielding a higher return than the childhood environment in the source country in general, with the specific result depending on the source country. The pattern could be due to a higher-quality education or health environment in the United States; yet at the same time, the quality of the health environment may not have been higher in the United States. For example, Eriksson and Niemesh (2016) show that children of black migrants during the Great Migration had higher infant mortality rates relative to children of non-migrants due to poor health conditions in northern cities. However, for states for which we have data at the turn of the twentieth century, infant mortality rates for foreign-born mothers were less than those for African Americans and similar to the rates in many European sources (Preston and Haines (1991), Tables 2.3 and 3.1). Ultimately, it is unclear whether migrating from Europe led to a lower quality health environment in the United States. Another possibility is that the U.S. environment was not better objectively, but rather that its education system trained individuals specifically for the U.S. labour market.

Despite the creation of the new linked sample of brothers, the age-at-arrival profile estimated without sibling fixed effects is within the standard errors of the profile estimated

[22] Standard errors are clustered at the household level.

Figure 2.3: The negative effect of age at arrival on the native-immigrant gap in wage income in 1940



Notes: The dependent variable is the age-adjusted gap in years of education between immigrants and natives. The figure shows the estimated fixed effects for age at arrival with age at arrival of zero and one being the excluded group. The shaded area is the 95 per cent confidence interval. Standard errors are clustered at the household level.
Sources: Linked sample of brothers from Ellis Island records to the 1940 Census.

with sibling fixed effects.[23] This result suggests that family-invariant unobservables such as parental preferences and education do not strongly bias the age-at-arrival profile estimated with OLS. However, it is also possible that we do not detect a difference in profiles between the methodologies due to errors in the linking process, which would bias the sibling fixed effects result towards the OLS result. This is likely not the case since we find the same result when limiting the sample to higher-quality links. In Appendix Table A.2, we keep the top 50 per cent of links in terms of quality of match based on closeness of name and year of birth and show that even with higher quality links, the sibling fixed effects method estimates a profile that is within the 95 per cent confidence interval of

[23]When not using sibling fixed effects, we control for country of birth and years in the United States.

the profile estimated with our main sample. We also show that if one links immigrants in a method related to Feigenbaum (2016), then the estimated profile is also within the confidence interval of the profile estimated with our main sample (see Appendix A.5 for more detail).

The negatively-sloped age-at-arrival profile appears to be linear; however, standard errors are wide so the profile may not truly be linear. Nevertheless, if one models the age-at-arrival effect to be constant across ages, then the effect of arriving one year later leads to a 1.1 per cent more negative wage gap with white natives. A linear age-at-arrival profile would go against expectations in two ways: first, others have found a steepening of the profile around the ages of 8 to 11 due to critical periods of language acquisition or health development (Bleakley and Chin (2004); Van den Berg et al. (2014)). Our estimate does show a dip in income between age 8 and 14, consistent with a critical period effect, but we cannot statistically detect a break in the slope. We also do not detect a steeper slope for ages under five, which may be surprising given the large returns to improved childhood environment during very young ages Almond et al. (2017); yet this may reflect the countervailing effects of a lower quality health environment in the United States relative to Europe.

Not only did arriving at an older age cause the native-immigrant gap to be more negative, but it also caused immigrants to enter lower skilled occupations relative to natives.[24] Table 2.3 shows that arriving at an older age increased the likelihood of entering an unskilled job and lowered the likelihood of holding a white-collar job. To provide a summary measure of the effect of age at arrival on occupation, the last two columns estimate the effect on occupational score and show that the native-immigrant gap for 16-year-old arrivals was five to 12 log points more negative than for infant arrivals.[25] Since

[24]The occupational categories are split by occ1950 codes such that professionals (codes starting with 0), managers (1), salesmen (3), and clerical workers (4) are white-collars. Farm owners, tenants, and managers (1) are farmers. Craftsmen (5) are skilled workers. Operatives (6), low-skilled service workers (7), farm laborers (8), and laborers (9) are unskilled workers.

[25]We show results based on a created occupational score that reflects mean earnings by occupation and source country in the 1940 Census. Creation of this score is discussed in Appendix A.4 and is largely based on Collins and Wanamaker (2017). We also show the age-at-arrival effect for the 1950 IPUMS variable occscore, the main one presented by Abramitzky et al. (2014). Results across scores differ because the 1940 score reflects a less compressed wage distribution and more adequately reflects immigrant earnings by occupation.

the magnitude of the effect on occupation score is less than the magnitude of the effect on income (17 log points), this suggests that age at arrival affected both occupation and income within occupation. Given these effects of age at arrival on income and occupation, it may be that age at arrival also affected other dimensions such as labour supply, but as shown in Appendix Table A.3, we find no effect of age at arrival on labour force participation or weeks worked.[26]

Table 2.3: Effect of age at arrival on occupations

Age at arrival	White-Col.	Skilled	Farmer	Unskilled	Log (Occ. Score) 1940	1950
2 to 3	–0.0198	0.00332	–0.0155**	0.0320	–0.0275**	–0.0103
	(0.0174)	(0.0170)	(0.00679)	(0.0196)	(0.0118)	(0.0137)
4 to 5	–0.00308	2.55e-05	–0.0176**	0.0206	–0.0478***	–0.0108
	(0.0175)	(0.0167)	(0.00697)	(0.0196)	(0.0120)	(0.0139)
6 to 7	–0.00276	–0.00804	–0.0317***	0.0425**	–0.0681***	–0.0140
	(0.0180)	(0.0174)	(0.00726)	(0.0203)	(0.0123)	(0.0142)
8 to 9	–0.0220	0.00812	–0.0332***	0.0472**	–0.0790***	–0.0175
	(0.0181)	(0.0175)	(0.00740)	(0.0203)	(0.0123)	(0.0144)
10 to 11	–0.0339*	–0.00117	–0.0379***	0.0730***	–0.0895***	–0.0288*
	(0.0190)	(0.0184)	(0.00776)	(0.0215)	(0.0131)	(0.0151)
12 to 13	–0.0332	0.000934	–0.0430***	0.0754***	–0.112***	–0.0383**
	(0.0211)	(0.0207)	(0.00873)	(0.0242)	(0.0144)	(0.0168)
14 to 15	–0.0720***	0.0184	–0.0428***	0.0963***	–0.113***	–0.0453***
	(0.0216)	(0.0213)	(0.00930)	(0.0246)	(0.0146)	(0.0173)
16 to 17	–0.0558**	0.00519	–0.0597***	0.110***	–0.122***	–0.0470**
	(0.0227)	(0.0227)	(0.00976)	(0.0260)	(0.0154)	(0.0183)
18 to 20	–0.0590**	0.0164	–0.0542***	0.0968***	–0.118***	–0.0437**
	(0.0244)	(0.0244)	(0.0112)	(0.0279)	(0.0167)	(0.0199)
N	49,904	49,904	49,904	49,904	49,904	49,904
R^2	0.545	0.511	0.603	0.550	0.700	0.557

Notes: *** p<0.01, ** p<0.05, * p<0.10.
The occupational score in the second to last column is the logged occupation based on mean wages in the 1940 Census (see Appendix A.4). The last column is the log occupational score based on the variable occscore in IPUMS. The excluded group is arrivals aged zero and one. Brothers fixed effects are included are clustered by household.
Sources: Linked sample of brothers from Ellis Island records to the 1940 Census.

We check the robustness of the age-at-arrival profile when controlling for observed birth order. Recall that observed birth order may not reflect the true birth order since we

[26] We also show in Appendix Table A.4 that age at arrival has no qualitative effect on home ownership and a small positive effect on living in a more urbanised location. Migrants were more likely to live in an urban environment, hence being consistently less likely to be engaged in farming.

do not observe family members left behind in the source country. In each specification, the age-at-arrival profile is unchanged, and the birth order effects are statistically insignificant. Another concern is that including immigrants who arrived aged older than 16 may bias results since older arrivals may have decided on their own to immigrate while younger arrivals had less choice. To account for this, we re-estimate the age-at-arrival effect when dropping those who arrived older than age 15, and find no difference in the estimated income profile. Finally, we test for the robustness of our linking process to the Americanisation process by relinking our data without Americanising names. The estimated effects without the Americanisation process have less precision (see Appendix Table A.7) but reaffirm the negative effect of age at arrival on adult labour market outcomes. These robustness checks are shown in Appendix Tables A.5–A.7.

2.6 Potential Mechanisms For The Age-at-arrival Effect

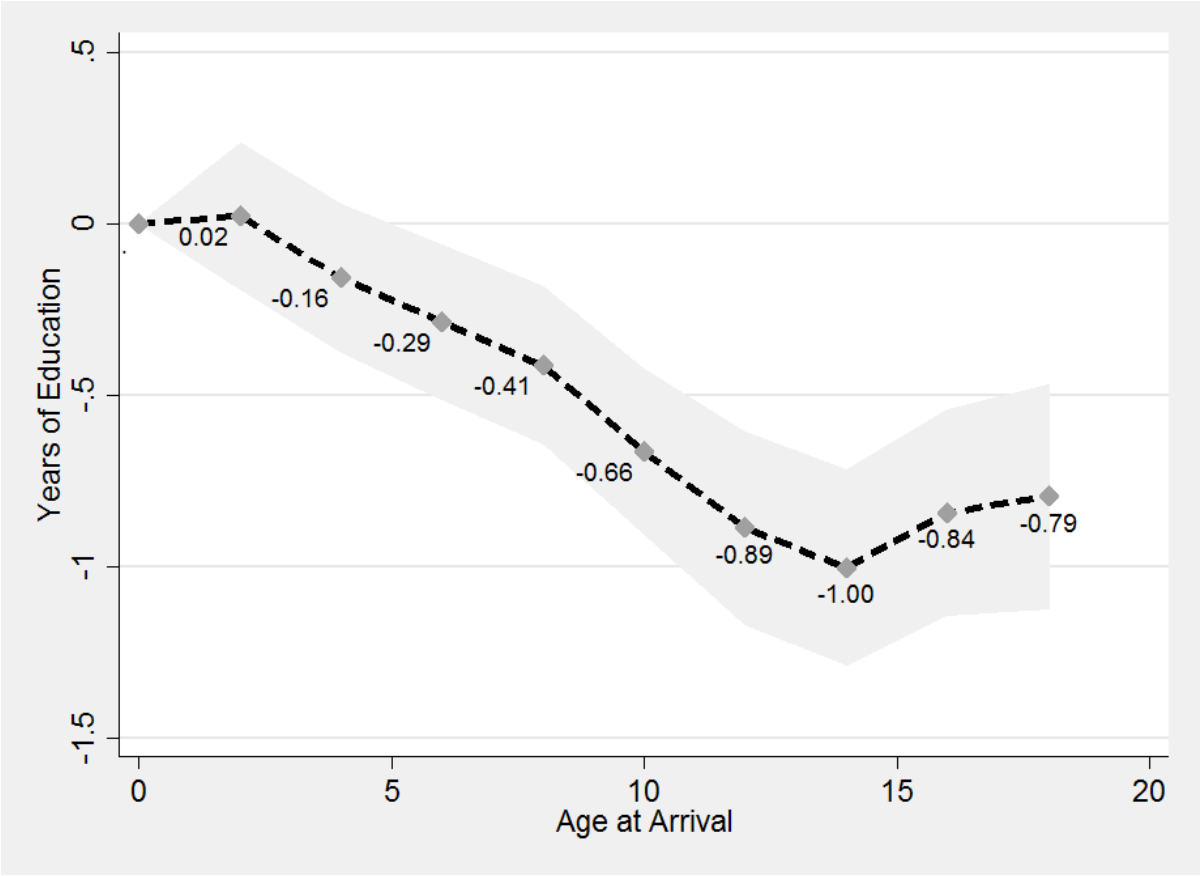
Arriving at an older age negatively affected labour market outcomes later in life, but through which channels? The potential mechanisms are numerous, but to name a few: the system of education changed across countries; parental resources may have improved after the move and thus investment into the child also improved; and younger arrivals may have socially adapted at a quicker rate. In this section, we estimate how age at arrival was related to these potential mechanisms.

2.6.1 Total Years of Education

First, we test whether age at arrival affected the total years of educational attainment. When we run the same age-at-arrival regression with native-immigrant gap in education as the dependent variable, we find that the gap for older arrivals was larger than the gap for younger arrivals by one year (see Figure 2.4). The education profile looks similar to the income profile in that they are both negatively sloped, suggesting that education could be a primary channel for the age-at-arrival effect on income. However, in contrast

with the income profile, the education profile becomes flat after age 15. The flattening of the education profile reflects that most immigrants left school prior to age 16 whether in the United States or in the source country.

Figure 2.4: The negative effect of age at arrival on the native-immigrant gap in years of education in 1940



Notes: The dependent variable is the age-adjusted gap in years of education between immigrants and natives. The figure shows the estimated fixed effects for age at arrival with age at arrival of zero and one being the excluded group. The shaded area is the 95 percent confidence interval with sibling effects. Standard errors are clustered at the household level.
Sources: Linked sample of brothers from Ellis Island records to the 1940 Census.

The effect of age at arrival on educational attainment may not have been the same across all source countries. Many Europeans arrived from countries with relatively robust education systems; for example, Germany had compulsory schooling laws dating back to 1717 and one of the best educational systems in Europe (Lindert, 2004). On the other hand, many Southern and European sources had less robust education systems; for instance, Italy and Greece had lower enrolment rates for five to 14 year olds compared

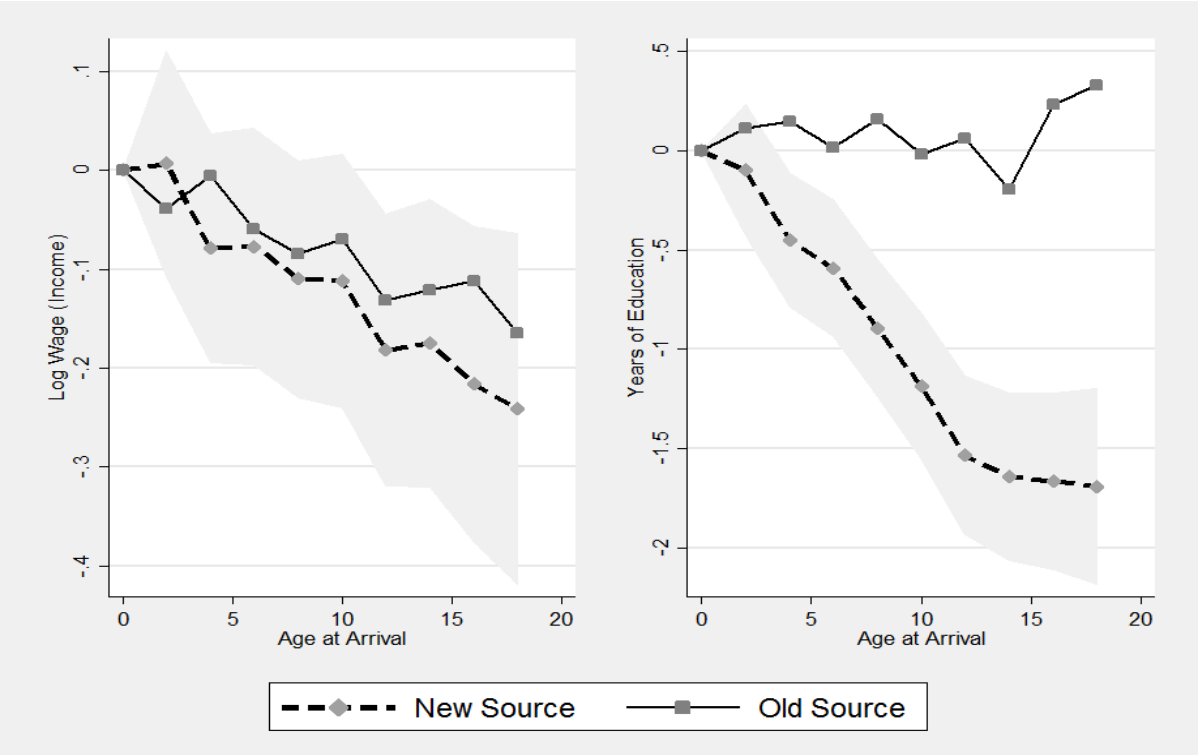
with the enrolment rates in Norway, Ireland, and the Netherlands (Bandiera et al. (2016), Figure 1). A reasonable hypothesis is that the effect of age at arrival on education is smaller for higher income countries in Northern and Western Europe compared with lower income countries in Southern and Eastern Europe.[27]

The profiles for income and education separated by New and Old sources are shown in Figure 2.5. On the one hand, the income profiles are similar across sources, where the negative effect of age at arrival is statistically indistinguishable across the two regions. At the same time, the education profiles were quite distinct: New sources had a steep profile where older arrivals received 1.7 fewer years of education. On the other hand, the education profile is completely flat for Old sources, showing no penalty for arriving at an older age. The flat education profile is consistent with the relatively high-quality educational institutions in Northern and Western Europe.

The difference in education and income profiles across sources reveal a puzzle: why did older arrivals from Northern and Western Europe earn less despite receiving the same total years of education? One reason may be that foreign education did not yield a high return in the U.S. labour market, and thus extra schooling acquired in the source country at older ages did not boost wages. Besides education, older arrivals also had more potential labour market experience in the foreign source country, and foreign experience may not have had a high value in the United States. On the other hand, it may be that the age-at-arrival effect operated through channels other than education or experience, such as social assimilation. We now turn to other possible explanations for the effect of age at arrival.

[27]We define 'Old Source' countries, or those from Northern and Western Europe, to be Denmark, Finland, Norway, Sweden, England, Scotland, Ireland, Belgium, France, Netherlands, Switzerland, and Germany. We define 'New Source' countries, or those from Southern and Eastern Europe, to be Greece, Italy, Portugal, Spain, Austria, Czechoslovakia, Hungary, Poland, Romania, Yugoslavia, Lithuania, and Russia.

Figure 2.5: The age-at-arrival profiles were differently sloped across new and old sources



Notes: The dependent variable is the age-adjusted gap in log wage income between immigrants and natives. Self-employed workers are dropped. The figure shows the estimated fixed effects for age at arrival with age at arrival of zero and one being the excluded group. The shaded area is the 95 per cent confidence interval when using sibling fixed effects. Standard errors are clustered at the household level.
Sources: Sample of brothers linked from Ellis Island records to the 1940 Census.

2.6.2 Return to Education and Experience Separated into Domestic and Foreign Component

Since we observe age of arrival and completed years of education in our dataset, we can test—after making a few assumptions—whether foreign education and experience yielded a small return in the United States. Following Friedberg (2000), we assume that individuals entered school at age six, and then allocate the total amount of schooling to the United States or source country based on the age of arrival.[28] Given this assumption, it is

[28]Let Total Education = Foreign Education + U.S. Education and Experience = Foreign Experience + U.S. Experience. Further assume that Total Experience = Age – Education – six. To separate total education and total experience into United States and foreign components, we assume that individuals attended schooling continuously. That is, let Foreign Education = Zero if Age at Arrival is less than six, and min(Age at Arrival – six, Education) if greater than or equal to six. Also let Foreign Experience = Zero if Age at Arrival is less than six, and max(Age at Arrival – Foreign Education – six, zero) if greater than or equal to six.

straightforward to further separate potential experience into foreign or U.S. components. See the descriptive statistics in Table 2.2 for how foreign education and experience increase at higher ages of arrival, while years of U.S. education decreases.

To measure the wage return to education and experience, we use an augmented Mincer equation and regress log income on years of U.S. education, years of foreign education, potential years of U.S. experience, and potential years of foreign experience:[29]

$$y_{ih} = \beta_0 + \beta_1 \text{ForEduc}_{ih} + \beta_2 \text{USEduc}_{ih} + f\text{ForExp}_{ih} + g\text{USExp}_{ih} + \sigma_h + \upsilon_{ih}.$$

We are interested in whether the return to foreign education was less than the return to U.S. education, that is if $\beta_1 < \beta_2$. We are also interested in whether the return to experience gained abroad yielded a different return from experience gained in the United States; we model experience as a quadratic. Note that we always include household fixed effects, eliminating household-invariant unobservables that are correlated with years of education, experience, and income.

The results are presented in Table 2.4. When pooling New and Old source countries together, the return to being educated in the United States is estimated at 5.3 per cent, which is less than the return for native-born workers (6.5–7.9 per cent). A different return to U.S. education for foreign-born workers relative to native workers has been found elsewhere in the literature (e.g., Chiswick (1978); Baker and Benjamin (1994)), and may reflect discrimination against foreign-born workers or the quality of "years" of education. The return to education earned in the foreign country was even lower at 4.4 per cent, although the 0.9 percentage point difference from the return to U.S. education is not statistically significant.

While there is not strong evidence that the location of schooling mattered, there is evidence that where one gained labour market experience mattered. Table 2.4 shows that the return to potential foreign experience was not statistically distinguishable from zero, although this is not very precisely estimated, whereas U.S. experience was positively

[29] There are potential endogeneity issues here as unobserved variables such as unmeasured or inherent ability, or motivational factors, may affect both income and schooling.

Table 2.4: The return to education and experience separated by U.S. and foreign component

Sample:	Full Sample	Only NW Europe	Only SE Europe	Full Sample
U.S. education	0.0528***	0.0645***	0.0415***	0.0645***
	(0.00692)	(0.00917)	(0.0105)	(0.00957)
U.S. educ.				–0.0233*
×SE Europe				(0.0138)
Foreign education	0.0439***	0.0583***	0.0347***	0.0583***
	(0.00535)	(0.00758)	(0.00772)	(0.00793)
Foreign educ.				–0.0236**
×SE Europe				(0.0108)
Foreign experience	0.00534	0.0109	–0.000811	0.0107
	(0.0125)	(0.0208)	(0.0164)	(0.0218)
Foreign exp.				–0.0113
×SE Europe				(0.0269)
(Foreign experience /10)2	–0.0130	–0.0712	0.00117	–0.0710
	(0.124)	(0.261)	(0.154)	(0.274)
(Foreign exp./10)2				0.0690
×SE Europe				(0.311)
U.S. experience	0.0686***	0.0796***	0.0564***	0.0795***
	(0.00889)	(0.0107)	(0.0150)	(0.0112)
U.S. experience				-0.0232
×SE Europe				(0.0181)
(U.S. experience /10)2	–0.117***	–0.143***	–0.0956***	–0.143***
	(0.0152)	(0.0192)	(0.0245)	(0.0201)
(U.S. exp./10)2				0.0469
×SE Europe				(0.0309)
Observations	35,229	15,881	19,348	35,229
R-squared	0.678	0.663	0.675	0.679

Notes: *** p<0.01, ** p<0.05, * p<0.10.
The dependent variable is log wage income. We assume that individuals enter school at age six and stay in school continuously in order to separate totals years of education and potential experience into foreign and U.S. components. Brothers fixed effects are included in each column.
Sources: Linked sample of brothers from Ellis Island records to the 1940 Census.

rewarded. A small return for foreign experience may reflect that immigrants entered different industries and occupations after the move to the United States. Immigrants often came from more agrarian countries in Southern and Eastern Europe, and their European experience appears to have had little value in the United States (Hatton and Williamson (1998), Chapter 2). However, this is part speculation, unfortunately we cannot observe the job history of migrants to determine the value of the type of foreign experience. Data

that does observe young arrivals' occupations show that most teenagers reported holding either no job or an unskilled job.[30] If skills learned in these jobs in the source country did not transfer to the U.S. labour market, then older arrivals would be penalised for staying longer in the source country.

The return to human capital may have differed between sources closer in development to the United States and sources further behind. We test for this in Columns 2 and 3 after splitting the sample into New and Old sources. The results show that the return to foreign education was indeed higher for Old sources at 6.5 per cent than for New sources at 4.1 per cent—a statistically significant difference, as shown in the fully interacted model in the last column. This result implies that education acquired in Northern and Western Europe more easily transferred to the United States, perhaps because it was of higher quality or the economies were closer in industrial structure. Yet at the same time, immigrants from Northern and Western Europe did not earn much return to foreign experience, and the return to foreign experience was similar across regions in Europe.

The results from these wage regressions point to foreign experience and its lack of return as a potential mechanism for a downward sloping age-at-arrival and income profile. For example, 16–20 year old arrivals had on average 5.7 years of potential foreign experience (see Table 2.2), but this human capital yielded little return in the U.S. labour market. This result may partially explain the result in Figure 2.5 that Old source immigrants had a negative age-at-arrival effect on income despite a lack of effect on education; however, other channels, such as the extent of social assimilation, may have also been important.

2.6.3 Social Assimilation: Intermarriage and Geography

Besides the traditional measures of human capital such as education and experience, age at arrival may have affected adult earnings through a different channel: social assimilation. This could be due to higher levels of English fluency for younger arrivals or because

[30] We can observe occupations using data from a random sample of ship arrivals to Ellis Island between 1917 and 1924, where the sample is limited to 12–17 year old males (Ward, 2017). According to this data, most arrivals reported having no occupation (40.6 per cent), with labourer (26.9 per cent), farm labourer (7.6 per cent), and farmer (3.4 per cent) being the top three reported occupations.

younger arrivals appeared more "American" and thus experienced less discrimination. We measure the effect of age at arrival on social assimilation by changing the dependent variable to outcomes related to residential segregation and marriage; specifically, the likelihood of living near native-born households and the likelihood of marrying a native-born spouse.[31] English fluency is not observed in the 1940 Census, but we discuss the potential effect of English fluency in the next section.

Age at arrival had a strong effect on intermarriage, as shown in Table 2.5. A 16-year-old arrival was 37.2 percentage points less likely to marry a native-born spouse than a one-year-old arrival, a very large effect given that 69 per cent of infant arrivals married a native-born spouse. Although we do not estimate the return to intermarriage, others have shown with late-twentieth century data that intermarriage is associated with higher earnings (e.g., Meng and Gregory (2005)); therefore, it may be a mechanism for the downward sloping age-at-arrival income profile.

While there is a large effect of age at arrival on intermarriage, there is little evidence that younger arrivals were more spatially integrated with native-born household heads. Table 2.5 also shows the effect of age at arrival on the fraction of native-born household heads in the county of residence. Note that we use fraction of native-born household heads rather than fraction of all individuals in the county to ensure that native-born second-generation children in the home do not influence the estimate (i.e., "childrearing" bias). We find that age at arrival had no effect on living in a county with more native-born household heads. However, this county-level measure may mask segregation within a county. Given that we have the entire 1940 Census, we can further narrow the geography from county to the immediate neighbourhood, as proxied by the fraction of native-born household heads on the same census page (Logan and Parman, 2017). However, even using this measure suggests that age at arrival had little impact on spatial assimilation for our dataset of brothers.

[31] When estimating these equations, we do not do the two-step process of predicting residuals for immigrants based on the native life-cycle profile and then regressing these residuals on age at arrival. This is because there may not be a well-defined relationship between native's age and marrying a native-born spouse or having a native-born neighbour. Therefore, we present results based on the simple age-at-arrival effects on the levels of having a native-born spouse or native-born neighbour. However, the results are qualitatively the same if we use the residuals after predicting the lifecycle profile with natives.

Table 2.5: Effect of age at arrival on social outcomes

	Intermarriage		Spatial Assimilation	
	Native Spouse	Spouse from Different Source	Fraction of County Native HH	Fraction of page Native HH
Age at arrival:				
2 to 3	–0.0505*	–0.0501**	–0.00411	–0.00382
	(0.0272)	(0.0254)	(0.00436)	(0.00817)
4 to 5	–0.0953***	–0.0760***	–0.00366	–0.00937
	(0.0274)	(0.0254)	(0.00449)	(0.00826)
6 to 7	–0.107***	–0.0974***	–0.00286	–0.00480
	(0.0280)	(0.0262)	(0.00464)	(0.00872)
8 to 9	–0.172***	–0.161***	0.000386	–0.00620
	(0.0280)	(0.0262)	(0.00463)	(0.00861)
10 to 11	–0.238***	–0.225***	–0.00308	–0.0109
	(0.0288)	(0.0269)	(0.00488)	(0.00901)
12 to 13	–0.274***	–0.261***	–0.00160	–0.00186
	(0.0321)	(0.0304)	(0.00556)	(0.0105)
14 to 15	–0.320***	–0.304***	0.00411	–0.0185*
	(0.0323)	(0.0307)	(0.00558)	(0.0105)
16 to 17	–0.368***	–0.341***	0.00331	–0.00895
	(0.0339)	(0.0323)	(0.00593)	(0.0111)
18 to 20	–0.419***	–0.408***	0.0111*	–0.0151
	(0.0360)	(0.0348)	(0.00662)	(0.0121)
Observations	38,803	38,803	53,129	53,129
R-squared	0.661	0.673	0.600	0.578

Notes: *** p<0.01, ** p<0.05, * p<0.10.
For the first two columns, we only include people who are married. The first column regresses whether the spouse is native-born on age at arrival, and the second column regresses whether the spouse is from a different source country on age at arrival. The fraction of page that are native household heads is the census page, which reflects immediate neighbours. Brothers fixed effects are included in each column.
Sources: Linked sample of brothers from Ellis Island records to the 1940 Census.

2.6.4 Other Unobserved but Potential Channels: English Fluency and Parental Investment

An important indicator of social assimilation and human capital that is not included in the 1940 Census is English proficiency. It is certain that English proficiency was lower for older arrivals from non-English-speaking countries, given the robust evidence for the critical period of language acquisition from Bleakley and Chin (2004) and Bleakley and Chin (2010). An indirect way to uncover the effect of English skills on adult outcomes is

to test whether the age-at-arrival income profile steepens at older ages for non-English-speaking sources relative to English-speaking sources (see Bleakley and Chin (2004) for a further discussion). However, we do not find that the age-at-arrival income profile for non-English-speaking sources steepens relative to the profile for English-speaking sources after the critical period of language acquisition ends, which is consistent with the argument that acquiring English fluency was relatively unimportant for improving one's occupation in the early twentieth century compared with the late twentieth century (Ward, 2018). Yet standard errors are wide when splitting the sample into English-speaking and non-English-speaking sources (see Figure A.1). Therefore, while a lower level of English fluency for older arrivals may have contributed to the negatively sloped age-at-arrival and income profile, this mechanism cannot be conclusively confirmed.

Finally, it is also possible that family real wages increased during the move; if so, then younger arrivals may have benefitted by receiving more parental inputs during critical stages of development. An increase to family income almost certainly occurred after migration: Abramitzky et al. (2012) estimate the return to immigration at 70 per cent from Norway to the United States, and the return was probably even larger for immigrants from Southern and Eastern European sources. Williamson (1995) estimates that U.S. real wages were 67 per cent higher than in Great Britain in 1905, and more than three times higher than real wages in Italy. Therefore, an additional mechanism for the age-at-arrival profile is likely that household investment into children increased and the effectiveness of this investment was higher at younger ages, but we do not observe the change in real income before and after the move.

Overall, we interpret the age-at-arrival effect as the effect of changing various environmental attributes at critical stages of childhood development. On average, older arrivals were penalised relative to younger arrivals because they had fewer total years of education, less valuable labour market experience, and were less socially assimilated. While we cannot precisely show which mechanism was most important, we do show that the effect of age at arrival was large enough to erase the negative native-immigrant wage gap that older arrivals experienced.

2.7 Conclusion

Using a new dataset of brothers linked from Ellis Island records to the 1940 Census, we show that there was a large wage and educational return to arriving at a younger age in the United States. Spending one's childhood in the United States rather than in Europe significantly improved immigrants' long-run economic outcomes. The variation in immigrant outcomes based on their age at arrival complements prior research that finds that occupational-based earning differentials between migrants and natives were fixed throughout the life cycle. The difference in results suggests that while human capital acquired during childhood led to a large occupational return, the human capital acquired during adulthood after arrival did not (Abramitzky et al. (2014); Ward (2018)).

While the results suggest that the U.S. childhood environment was advantageous relative to that of Europe, the results are limited by the lack of data on childhood location in the United States. In particular, one question that is left unanswered is the effect of childhood environment when living in or outside an ethnic enclave. Given the intergenerational literature's result that the source country's position in the occupational distribution persists across generations of immigrants despite our result of childhood environment having a large effect, persistence of skill levels across generations must be due to other factors such as immigrants sorting into different quality childhood environments in the United States (Abramitzky et al. (2014); Borjas (1994)).

Finally, the results show that immigrants' position in the skill distribution was not fixed by inherited or genetic factors, as many nativists at the time claimed. For example, Francis Walker, one-time president of the American Economic Association, charged that New source immigrants had "none of the inherited instincts and tendencies which made it comparatively easy to deal with the immigration of the olden time. They are beaten men from beaten races; representing the worst failures in the struggle for existences. Centuries are against them" (Walker, 1896). This pessimistic view of immigrants has been reprised throughout time, right up to today's wave of nativism against Muslims and Hispanics (Higham (1955); Huntington (2004)). Our research shows that these "beaten men from beaten races" in the past did remarkably well with a change to their location at

young ages, reaffirming the paramount importance of childhood environment for long-run outcomes.

Chapter 3

The Making of a Nation: Who Voted for Australian Federation?

3.1 Introduction

An influential literature has analysed the size of nations and the economic and social foundations of their formation and dissolution. The political integration and disintegration of nations is seen as endogenously determined by the interaction between agents or coalitions pursuing objectives subject to constraints. This literature has emphasised the trade-off between the economic benefits of greater size and the economic costs of increasing diversity (Alesina and Spolaore, 2003). The benefits of merger or integration include spreading the cost of public goods over a larger tax base, and may also include (depending on the external trade regime) economies of scale in production and the gains from trade (Alesina et al., 2000). The costs of integration arise from increased bureaucracy or congestion but most importantly from greater diversity. While diversity may be a source of gains from complementarity or specialisation, it may also be a cost if preferences that differ between groups lead to distributional struggles or are otherwise difficult or costly to accommodate in a single set of laws or regulations (for a useful survey, see Spolaore (2016)).

As more than fifty new nations have emerged since 1946, most of the focus has been on dissolutions. And while the underlying theory, based on interest group preferences,

focuses on efficiency gains and redistribution, secessions often involve wars and external geopolitical forces (Bolton and Roland, 1997). The evidence suggests that civil wars and irredentist movements or secessions are related to ethnic polarisation and cultural heterogeneity (Montalvo and Reynal-Querol (2005); Desmet et al. (2011)). In contrast there are few studies within this framework of how and why mergers take place peacefully to form new nations through a democratic process.

In this paper we study one key example: the federation of six quasi-independent British colonies to form the Commonwealth of Australia in 1901. After at least a decade of debate over whether to unite (Brown, 2004), and under what terms, in 1898 four of the then colonies conducted referendums. The other two, Queensland and Western Australia, failed even to agree on having a referendum. The legislatures of New South Wales, Victoria and Tasmania imposed a minimum number of affirmative votes required to pass the respective bills into law. It failed to pass only in New South Wales, which set the highest bar. After further negotiation a second round of referendums took place in 1899/1900. In this second round of voting, federating was approved in all six colonies and the Commonwealth of Australia, in which the former colonies became states, came into effect on 1 January 1901.

The historical literature on Australian Federation has focused on a range of influences. Possible gains from increased scale include establishing a combined defence force and gaining better terms for infrastructure loans from the London capital market. But as the new Commonwealth of Australia remained closely tied to Britain such advantages would be marginal. A more important motive may have been to create a customs union, where previously each colony had a different tariff regime, which applied both externally and between colonies. A related issue was to enhance commerce through integration of transport infrastructure. On the cost side, ethnic diversity was hardly an issue in a setting where the overwhelming majority of the (white) population had British or Irish ancestry. But there were some lower-level cleavages by religion, gender and age, as well as differences in attitude between those that had moved into the colony and those that were born there. Geographical location was also potentially important, particularly at the

borders between colonies, where fiscal and administrative barriers would be more evident, and in areas remote from the metropolitan centres.

In this paper we analyse the patterns of voting in the referendums to tease out some of the possible reasons to vote for federating. We relate the share of support for federating to the characteristics of districts as observed in the 1901 Colonial Censuses. This is made difficult because of the mismatch between electoral districts and census districts. In order to overcome this problem, we have obtained the votes for each polling station, which we have geocoded and then re-aggregated into census districts.

The results of estimating across districts indicate that economic interests, represented by broad economic sectors, are not correlated with support for federating in the manner predicted. Religion and age structure also do not seem to be correlated. The strongest associations with affirmative votes are the share of immigrants (international or inter-provincial), the distance from the metropolitan centres, and the share of females in a district. To economists it may seem surprising that economic interests do not influence support for federating in the way that the theory of customs unions would suggest; perhaps voters have other motivations. On the other hand, it is consistent with the finding that there is little evidence for ex-post trade creation (Irwin, 2006). One caution with our results is that they are associations, not causal, and it is important to recognise that there is a difference between the timing of our dependent variables (1898, 1899, and in one case 1900), compared with our explanatory variables, which are mostly as at 1901.

3.2 Background to Federation

European settlement of Australia began in 1788 with the arrival of the First Fleet bringing convicts from Britain. By the 1830s free settlers were arriving to take advantage of the opportunities in the colony of New South Wales, which by 1827 embraced the eastern two-thirds of the continent of Australia plus what is now New Zealand. In the ensuing decades parts of this vast area were detached to form separate colonies under the British

Crown.[1] From the 1850s when legislatures were first established these colonies evolved distinct identities and divergent policies, although, as British colonies, their administrative structures retained much in common.[2] As separate colonies, they adopted different railway gauges, they evolved different fiscal arrangements and they developed different tariff policies, both for external and inter-colonial trade.

From mid-century onwards there were periodic discussions and proposals for reuniting the colonies as one Dominion embracing the whole continent.[3] 1885 saw the establishment of a Federal Council at which representatives of the colonies met every two years to discuss issues of common interest. But it had no executive power and it was a forum for cooperation rather than a step towards federating. The movement for federating that would be successful was launched by Sir Henry Parkes (often referred to as the Father of Federation) in a speech at Tenterfield (NSW) in 1889. Although defence was the initial focus,[4] subsequent speeches by Parkes and other politicians, included issues such as the control of non-white immigration, the benefits (and costs) of inter-colonial free trade and external tariff reduction, as well as the unification of the railways and the regulation of water rights. Parkes initiated a conference of colonial Premiers in Melbourne in 1890 and one in Sydney, but little progress was made.

Meanwhile several federation leagues became active, and those in the border districts of New South Wales and Victoria convened a conference at Corowa (NSW) in 1893. The Corowa conference included 74 delegates from associations representing a range of social strata. Under the leadership of Dr John Quick, it unanimously proposed that another constitutional convention be organised with delegates elected by the people. It further proposed that the constitution drafted at this convention be put to the people

[1] In 1827 Van Diemen's Land was proclaimed as a separate colony, renamed Tasmania in 1856, and in 1829 the western third of the continent became the Swan River Colony, renamed Western Australia in 1832. This was followed by the creation of other colonies from parts of New South Wales: South Australia in 1836, Victoria in 1851 and Queensland in 1859. New Zealand, loosely attached to New South Wales, became a separate colony in 1841.

[2] Queensland gained self-government in 1867 and Western Australia not until 1890.

[3] For contemporary commentaries on the political debates and negotiations leading up to federating, see Quick and Garran (1901), Wise (1913), and Deakin (1944).

[4] The original pretext was a report by British Major General Bevan Edwards, written in the wake of incursions by France in New Caledonia and by Germany in New Guinea, which stated that the separate colonial defence forces were inadequate.

in a referendum. This idea was approved by the colonial premiers who, at a meeting in Hobart in 1895, produced a Draft Enabling Bill based on the Corowa proposals and consisting of 39 sections. Among the issues current at that time were trade, tariffs and the financing of a federal government. Some of these were discussed at a people's conference in Bathurst (NSW) in 1896, in particular the redistribution to states of federal customs revenue. This was followed a constitutional convention which met three times in 1897–8, which was attended by representatives from each colony and which debated and refined a Bill to place before the electorate for approval.[5]

The first round of referendums took place in four colonies in early June 1898. These were organised on the same electoral districts as for elections to the lower houses of the colonial legislatures.[6] But the issue was decided by an overall majority of all votes in the colony. However, New South Wales, Victoria and Tasmania also imposed a minimum number of affirmative votes required to pass the Bill. This was highest in New South Wales, which laid down a minimum of 80,000 votes in the affirmative in order to pass the Bill. The minimum in Victoria was 50,000 and in Tasmania 6,000. The franchise was similar to that for parliamentary elections but it differed between colonies.[7] In South Australia it was all adults aged 21 and over (women having been enfranchised in 1894), but in New South Wales and Victoria, it was restricted to adult males and in Tasmania there was an additional property qualification. As Table 3.1 shows, in the four colonies that voted in the referendums of 1898, the total number on the electoral roll was 727,438, or 23.8 per cent of their total population.

The turnout in the 1898 referendums was less than half in each of colonies that voted, and it amounted to 45.5 per cent overall. Of those that did vote, the overall percentage of support for federating was close to two-thirds. In Victoria it was 81.3 per cent, in Tasmania 79.9 per cent and in South Australia 66.5 per cent. But in New South Wales only 51.2

[5]The meetings took place in Adelaide in March 1897, in Sydney in August 1897 and in Melbourne in January 1898. They were attended by ten elected delegates from each colony except Queensland, which did not participate.

[6]In New South Wales and Western Australia there were single member constituencies while the other colonies had multi-member constituencies, which were therefore somewhat larger on average.

[7]In Victoria, Western Australia, South Australia and Tasmania there was a property qualification for voting in the parliamentary elections (Rhodes, 2002, p. 10).

Table 3.1: Electors and voting at the 1898 and 1899-1900 referendums.

	NSW	Vic	QLD	WA	SA	Tas
Referendums in 1898						
Vote date	1898-06-03	1898-06-03	–	–	1898-06-04	1898-06-03
Population	1,346,240	1,175,463	498,523	168,128	362,897	177,340
Electoral roll	306,878	252,560	–	–	136,387	31,613
Votes cast	138,657	123,627	–	–	53,836	14,697
Turnout (%)	45.2	48.9	–	–	39.5	46.5
Yes votes	71,595	100,520	–	–	35,800	11,746
No votes	66,228	22,099	–	–	17,320	2,689
Yes majority	5,367	78,421	–	–	18,480	9,057
Referendums in 1899-1900						
Vote date	1899-06-20	1899-07-27	1899-09-02	1900-07-31	1899-04-29	1899-07-27
Population	1,357,050	1,176,854	512,541	179,022	365,755	182,508
Electoral roll	307,473	287,331	107,133	89,593	152,554	34,528
Votes cast	191,327	163,783	69,832	65,030	93,952	14,342
Turnout (%)	62.2	57.0	65.2	72.6	61.6	41.5
Yes votes	107,420	152,653	38,488	44,800	65,990	13,437
No votes	82,741	9,805	30,996	19,691	17,053	791
Majority	24,679	142,848	7,492	25,109	48,937	12,646

Source: Rhodes (2002, pp. 12, 14 and 16).

per cent voted in favour and the total number of supportive votes fell short of the 80,000 threshold. Thus the first round of referendums failed to provide unanimous support for federating and the other three colonies that voted were not prepared to proceed without New South Wales. Queensland failed to have a referendum in 1898 partly because of changes in leadership and partly because of differences of opinion over whether the colony should be divided into three electorates (north central and south), and require majorities in all three in order to approve the draft constitution. In light of the result in New South Wales, Western Australia, where opinion was also deeply divided, decided not to proceed with a referendum (de Garis, 1999, p. 303).[8] So the federation process ground to a halt.

Following the failure to reach the minimum yes vote, the New South Wales Premier (George Reid) met with the other premiers to discuss amendments to the draft constitution and upon agreement to propose another round of referendums. These amendments included the requirement of a two-thirds majority for a joint sitting of both Houses following a double dissolution and the provision that the Federal Capital be in New South

[8] In any case the Enabling Acts of Queensland and Western Australia contained provisions that they would not join a federation that did not include New South Wales (Rhodes, 2002, p. 10).

Wales (but at least 100 miles from Sydney). They also included placing a time limit of ten years on the scheme for redistributing revenue among the states and variations to the process for altering the constitution. Once these amendments were agreed New South Wales resolved that the referendum would be decided on a simple majority, with no minimum threshold. The agreement on the second round of referendums also included the rider that New South Wales would vote first and that the other colonies would follow in the light of that result.

In the event South Australia voted first (because of slow progress in New South Wales), and the other colonies except Western Australia held their referendums within a few months of New South Wales. Western Australia, where the politicians were more divided, delayed introducing an enabling bill for a referendum until all the other colonies had voted.[9] In the second round of referendums turnout was considerably higher in three of the colonies that had voted in the first round but lower in Tasmania. The overall turnout of 61.1 per cent reflected the increased public salience of the federation debate, especially among the middle class.[10] In the two more reluctant colonies, turnout was particularly high but the majority in favour was smaller in Queensland than in Western Australia. Among the nearly 600,000 votes cast, 70.7 per cent were in favour of federating. But there were sharp differences, with Victoria and Tasmania recording over 90 per cent in favour while in New South Wales and Queensland the percentages were 56.1 per cent and 55.1 per cent respectively. Nevertheless, the affirmative vote in New South Wales easily exceeded the discarded threshold of 80,000. In Queensland and Western Australia, where politicians wrestled extensively over federating, the affirmative vote exceeded two thirds.

Following the 1899 referendums a deputation of representatives from each of the colonies travelled to London to lobby the British government (in particular the Colonial Secretary, Joseph Chamberlain) to approve the Commonwealth Constitution Bill. After reaching a compromise over appeals to the Privy Council the Bill passed through

[9] One thing that brought Western Australia's dithering over the referendum to a head was the threat of the eastern goldfields areas to secede from the colony in order to join the Commonwealth (de Garis, 1999, p. 311).

[10] Analysing turnout at the individual level for Bendigo (Victoria) in 1899, Fowler (2013) finds that property owners were ten percentage points more likely to vote than non-owners.

the Westminster parliament and received royal assent on 9 July 1900. Western Australia, which had delayed for further concessions had not yet voted although a date had been set. Three weeks after the approval in London, Western Australia voted in favour of joining on a franchise that, for the first time, included women. On 1 January 1901 the Commonwealth of Australia came into effect as a federal system, including all six former colonies as states and encompassing the whole continent.

3.3 Debates and hypotheses

Our paper aims to provide a better understanding of who voted for the Australian Federation and why. The theory of customs unions suggests that areas dominated by economic sectors that would benefit from trade, such as agriculture, would be more supportive. The historical debate initiated by Parker (1949) and Blainey (1950) focused on the possible economic imperatives. But later contributions shifted towards politics and popular culture as the key influences. And they focused more on regional and local developments, drifting away from the bigger picture. With a few partial exceptions there has been no comprehensive quantitative analysis of voting patterns in the referendums leading up to federating.[11]

In terms of the gains from greater economic scale, one might expect the most populous colony, New South Wales, to be the least favourable to federating, as proved to be the case, but Victoria, which was almost as large, was strongly in favour. Defence and immigration policy, often mentioned as issues early in the campaign, rapidly faded from the debate, except in Queensland and to a lesser extent Western Australia, although it remained discussed in newspapers, for instance Alien Immigration (1898).[12] One reason is that a united Australia would remain firmly in the British empire and would look to Britain for both leadership and in foreign policy and material assistance in defence.[13]

[11] These exceptions are Rhodes (1988) who explored correlations between yes votes and the positions of local politicians and newspapers, and Coleman (2017) who estimates the relationship between yes votes and a range of socioeconomic variables for South Australia.

[12] The impression that (non-European) immigration was an issue is supported by the fact that one of the first pieces of legislation passed by the federal parliament was the Immigration Restriction Act 1901, which inaugurated the so-called White Australia Policy.

[13] Full independence from Britain was never an issue, except among a few radicals (Eddy, 1978a). It is

Perhaps a more important background factor was the severe recession of the 1890s, when high unemployment, bank failures and the drying up of British loans in 1894 concentrated minds on economic issues.[14] This helped to build support among politicians from across the political spectrum in favour of federating, some of whom thought that a united Australia would make it easier to borrow in London. This could help to explain why Victoria, the worst affected colony, would be overwhelmingly in favour of federating. Set against this were concerns that an additional layer of government would lead to higher taxes, either direct or indirect.

Initiating a now famous debate, Parker (1949) argued that the degree of support for federating depended on which economic sector dominated a particular region and whether or not its interests would be advanced by reduced barriers to trade within Australia as well as by the expected external tariff regime of a united Australia. He also pointed to higher levels of support in areas close to the borders between colonies, notably the Riverina districts on the border between NSW and Victoria. In his riposte to Parker, Blainey (1950) criticised the broad regional approach, suggesting instead that a wide range of social and political influences were at work and that their effects varied both within and between electoral districts, a point echoed by (Bastin, 1951, p. 205).[15][16] Nevertheless, Norris (1978, p. 192) concluded that "[b]y and large, attitudes to federation owed less to political persuasion than to expectation of economic gain or loss." The economic gains and losses would be those arising from a common external tariff, the abolition of inter-colonial customs duties, the implications of federal control over revenue, and the unified management of railways and water resources.

Tariff unification was an important issue because pre-federating tariffs, which applied both to external and inter-colonial trade, varied widely between colonies. As illustrated

notable that even a century later (in 1999) a referendum on moving to a republic failed to gain a majority (see McAllister (2001) for an outline of long term trends and an analysis of the referendum).

[14] Merrett (2013) discusses the banking crisis noting that colonial governments had very little capacity to assist banks in distress, noting also that Federation passed legal power over banking and finance to the Commonwealth.

[15] In his reply to Blainey, Parker (1950) largely conceded this point.

[16] Much of the subsequent literature focused on the debates within colonies and electorates. These include Hewett (1969) and Irving (1999), on New South Wales, Norris (1969), Pettman (1969) and Bannon (1999) on South Australia, and Hillman (1978) and de Garis (1999) on Western Australia.

Table 3.2: Estimates of Average Import Duties in 1900

	Per cent of imports on duty-free list	Per cent ad-valorem rate of duty	
		On dutiable merchandise imports	Excluding "narcotics and stimulants"
NSW	87.6	10.3	1.3
Vic	53.4	36.2	17.0
QLD	36.0	20.5	13.1
SA	35.7	21.8	14.0
WA	37.1	14.8	9.3
Tas	9.0	24.2	22.0

Source: (Irwin, 2006, p. 317).

in Table 3.2, tariffs were highest in Victoria (and covered a wider range of imports, particularly manufactured goods) and lowest in New South Wales. The other colonies had smaller percentages on the duty-free list and average tariff rates somewhere between New South Wales and Victoria. All of the colonies levied high tariff rates on alcohol and tobacco (classified as narcotics and stimulants) largely for revenue raising purposes.[17] When these items are excluded the average tariff rates on the remaining dutiable items are substantially lower, as illustrated in the last column of Table 3.2. In the more protectionist colonies the goods subject to duty were mainly manufactured and semi-manufactured goods, with rates of duty that varied widely across goods and between colonies (Lloyd, 2015, p. 170).[18]

Which sectoral interests would benefit from tariff reform would depend on expectations about the post-federating tariff structure. It seems reasonable at first sight to suppose that voters would have expected the unified tariff to resemble the Victorian tariff, which is roughly what emerged (Forster, 1977). One reason is that the tariff would be the main source of revenue for the federal government.[19] As the so-called Braddon clause provided that three-quarters of tariff revenue would be returned to the states, the room for a lower

[17] There were also excise taxes on these commodities but the rates were generally much lower than the tariff (Lloyd, 2017, p. 54).

[18] In the more protectionist colonies, most of the goods on the duty free list were intermediate inputs, raw materials and foodstuffs, but in the smaller colonies there were also substantial tariffs on imported foodstuffs such as wheat, oats, flour, eggs, meat and fruit (Lloyd, 2015, p. 171).

[19] Tariffs were an important part of colonial government revenue. In 1898 customs duties accounted for the following percentages of tax revenue and total revenue respectively: New South Wales: 63.7, 16.5; Queensland: 84.9, 35.2; South Australia: 68.2, 23.3; Tasmania: 79.1, 47.1; Victoria: 78.2, 30.6; Western Australia: 90.6, 36.4 (Barnard, 1985).

tariff was limited.[20] This clause, originated by Edward Braddon, Premier of Tasmania, was agreed in the Melbourne session of the constitutional convention in 1898, to resolve the contentious issue of revenue sharing. It was widely opposed in New South Wales where it became known as the 'Braddon Blot'. After the first round of referendums the Braddon Clause was time-limited to the first ten years. On the assumption of a high post-federating tariff, voters in manufacturing districts in New South Wales might be expected to support federating, although consumers would suffer from potential higher prices (trade diversion). In Victoria and the other colonies manufacturing districts might also be expected to support federating, not least because they could sell into the newly protected market of New South Wales (trade creation). In contrast, districts that specialised in export-oriented primary commodities would likely be opposed to federating. That all said, it is difficult to know what information the average voter possessed. While it is certainly the case that the federation issue was being covered in newspapers, there was a variety of opinions. For instance, one may think that Tasmania, as a poorer colony would have been supportive of federating, however at least amongst some there was a feeling that the colony was turning the financial situation around and that federating could make the situation worse (Federation and Taxation, 1899).

There was little ethnic diversity in Australia at the time of the referendums as the vast bulk of the (white) population had UK ancestry.[21] But interests could differ along other lines, one of which was religion. The 1901 censuses reported 39.7 per cent of the population as Anglican and 22.7 per cent as Catholic. While Catholics inspired by Irish republicanism might have supported federating as a step towards autonomy from Britain, some saw it as strengthening the Anglican ascendency (Cahill, 2001). Another possible fault line is gender. Women's associations generally supported federating, especially those seeking to gain female suffrage (Irving, 1999, Ch. 10). But women were entitled to vote only in South Australia and Western Australia and the draft constitution did not explicitly provide for

[20] Total tariff revenue would also shrink due to the loss of inter-colonial tariff revenue. This was a particular concern in Western Australia, which in 1900 negotiated that its tariff (inter-colonial and external) would remain unchanged for the first five years after Federation.

[21] Aborigines, who accounted for 2.4 per cent of the total Australian population in 1901, did not have the right to vote in the referendums, except in South Australia, and did not gain the federal franchise until 1962.

universal female suffrage although it may have been implied.[22] There was also potential for differences by education. Some writers point to a utopian vision of a united Australia as a growing element in the support for federating (Birrell (1995); Irving (1999); Martin (2001); cf. Atkinson (2013)). The Australian Natives Association, which was founded in Victoria, campaigned vigorously for federating (Blackton (1958); Pettman (1969)). It was led by a younger generation of politicians and, to the extent that its progressive ideals diffused into the general population, one might have expected them to have found greatest resonance among the younger and more educated voters.

Inter-colonial migration was also potentially important. Those living in the colony in which they were born might have a loyalty to their particular colony and a resistance to change that would not be shared by those born elsewhere. The most well-known example is the migrants from other colonies living in Western Australia (so called 't'othersiders), many of whom migrated in the 1890s gold rush and may have had a substantial influence in that colony (Bastin (1951); Hillman (1978)). Immigrants from overseas (mainly from Britain) may also have supported federating, as did the UK government.

Geographical differences are likely to have been more influential on voting on federating. Lack of unified management of railways and waterways often meant additional costs of portage or longer and more expensive routes to and from markets would have been felt most on the borders between colonies. One impediment was that lack of integration of the railways, where the colonies operated three different railway gauges.[23] Also, water transport was especially important in the Riverina districts on the border between New South Wales and Victoria. Interestingly this was where some of the most active campaigning for federating took place. Another hypothesis related to geography is that federating would favour places farthest from the metropolis in each colony. But voters in districts remote from the existing seats of government might support or oppose federation for other reasons

[22] The Constitution (Chapter I, Part IV) provided only that the franchise in federal elections would be the same as that already existing in each state, and it seems likely that having partial female suffrage would have been untenable. In any event, the franchise for federal elections was extended to all women aged 21 and over in the Commonwealth Franchise Act, 1902. For state elections, women were enfranchised in New South Wales in 1902, Tasmania in 1903, Queensland in 1905 and Victoria in 1908.

[23] There were three different railway gauges in common use. New South Wales was on standard gauge, 4 ft 8.5 in, Queensland, Tasmania and Western Australia adopted narrow gauge, 3 ft 6 in, Victoria chose broad gauge, 5 ft 3 in, and South Australia operated a combination of broad gauge and narrow gauge.

too. For instance, it is possible that those in the north of Queensland had no respect for colonial boundaries or loyalty to the colonial administration in Brisbane (Bolton, 1963, pp. 209–10).

Across districts, some of the influences identified in the literature (or proxies for them) can be captured with variables from the census but others cannot. Perhaps the most prominent omissions relate to politics and persuasion which came to the fore as the referendum campaigns intensified. The political strength of labour had grown with increasing representation in colonial legislatures, despite the decline in trade unionism since the wave of strikes and unrest of the early 1890s. The labour movement, which was barely represented at the constitutional convention, offered little support for federating as it was seen as serving conservative and middle class interests (Eddy, 1978b). In New South Wales labour opposed federating partly on the grounds that it would further weaken trade union bargaining power. Yet certain issues of interest to labour, such as federal control of pensions and the arbitration system were incorporated into the final draft as 'specific powers' of the Commonwealth (Martin, 2001).[24]

While politics and persuasion have been prominent in the literature, voting did not simply follow party lines. For instance, at the three meetings of the constitutional convention, voting was often divided across political affiliations and across representatives of different colonies (Loveday, 1972). In the 1898 referendum, in one-third of electoral districts the majority voted in the opposite direction to the position of their elected representatives (Rhodes, 1988). Indeed, among Labor members of the New South Wales Parliament 17 out of 19 were opposed yet 15 of their constituencies approved it. And in Tasmania where the Hobart newspapers were strongly anti-federation, the vote was overwhelmingly in favour (Warden, 1999, pp. 212–213).

[24] Another issue of interest to labour was to ban non-white immigration, something that came about in the one of the first acts of the Federal Parliament, the Immigration Restriction Act, 1901.

3.4 Data

In order to analyse voting patterns at the district level within each colony we need to match the data on voting patterns with explanatory variables from the census. Voting took place within electoral districts, which varied in size and were based on constituencies in the colonial legislatures. Unfortunately, for four of the six colonies these differ from the districts on which most variables are reported in the census. Rhodes (2002) produced a compendium of voting by electoral district, based on contemporary sources. These were compiled from data by polling station and we have obtained the underlying data on the votes cast, for and against, at each individual polling station. This means that we can reorganise the data on voting to match the geography of the census.

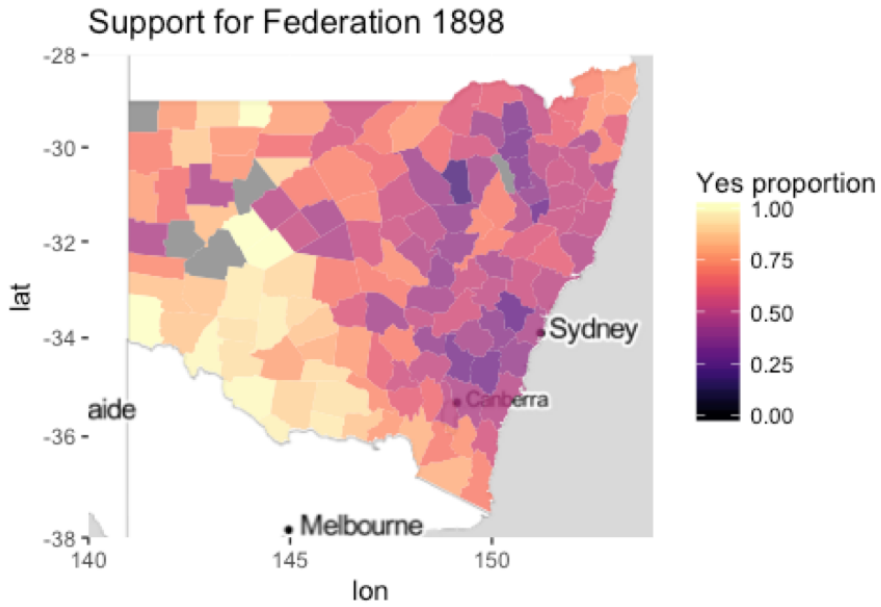
For our explanatory variables we use district-level variables from the 1901 censuses. Because there is very little surviving unit record data from the Australian censuses we have to rely on the published volumes. One concern is that almost all of the explanatory variables post-date the dependent variable. However, unless there were major changes in the population in just two or three years, using the 1901 censuses should be appropriate. We use the 1901 censuses rather than 1891 censuses because the latter predates the deep recession of the mid-1890s and, more importantly because we have maps of the relevant census district boundaries only for 1901. The districts for which most of our variables are reported differ between colonies. These are counties in New South Wales and Victoria, census districts in Queensland, magisterial districts in Western Australia and electoral districts in South Australia and Tasmania. In order to match the polling station results with the census units we first locate and geo-code all the individual polling stations within the electoral districts by identifying the locality or the specific place (such as a post office, hotel or homestead).[25] We then geo-code the boundaries of all the census districts using maps of the borders provided by Camm et al. (1983). Finally, we re-aggregate the polling station data into these districts.

To illustrate the outcomes, Figure 3.1 shows the share of yes votes by county for New

[25] These places were located in the first instance using the Google Maps API, and then supplemented using the websites `LatLong.net` and `Bonzle.com` to find the ones where there were alternatives with the same name or where the name is no longer used.

South Wales, which is the most populous colony, in the 1898 vote. Although 52 per cent of all formal votes were supportive, there was wide variation across districts. The yes vote was strongest in the districts to the south and west, particularly in the Riverina region, and it was weakest in some of the districts around Sydney, notably the Southern Highlands.

Figure 3.1: Share of 'yes' votes of all formal votes cast in NSW in 1898



Sources: Data from Rhodes (2002) and maps from Camm et al. (1983).

New South Wales had two votes. The first vote was on 3 March 1898, and the second vote was on 20 June 1899 (apart from the area of St Leonards, which is a suburb of Sydney, where it was held 10 days later on 30 June 1899). There were 137,738 formal votes cast in 1898, and this increased by around 37 per cent to 188,621 in 1899. Much of that increase was due to supporters of federating. 'Yes' votes increased by about 50 per cent, while 'No' votes only increased by about 25 per cent. In 1898 the census area that was least-supportive of federating was Baradine at 20 per cent, followed by Cook with 28 per cent, and Parry with 29 per cent. In 1898, the most-supportive districts were Tara, Thoulcanna, Werunda, and Woore, all four of which had every formal vote supportive of federating. However, those four areas had a low number of votes. Wakool was the most supportive of the districts with a significant population, at 99 per cent support. In 1899 support for federating in Baradine decreased to 15 per cent, but the next least

Table 3.3: Descriptive statistics

Variable	NSW		QLD		SA	
	Mean	SD	Mean	SD	Mean	SD
Proportion of male adults	0.60	0.08	0.60	0.08	0.52	0.08
Proportion adults aged 21–29	0.29	0.04	0.29	0.05	0.29	0.03
Proportion born in colony	0.67	0.19	0.53	0.13	0.75	0.13
Proportion primary industries	0.57	0.13	0.47	0.20	0.41	0.22
Proportion manufacturing	0.08	0.04	0.10	0.07	0.14	0.09
Proportion Catholic	0.29	0.07	0.26	0.07	0.14	0.06
Proportion literate	0.90	0.04	0.85	0.13	0.93	0.05
Proportion border areas	0.25	0.43	0.17	0.38	0.26	0.45
Average distance to capital (km)	469	227	675	598	171	286
Number of districts	136		63		27	
	TAS		VIC		WA	
	Mean	SD	Mean	SD	Mean	SD
Proportion of male adults	0.52	0.04	0.54	0.03	0.69	0.12
Proportion adults aged 21–29	0.30	0.04	0.29	0.02	0.32	0.06
Proportion born in colony	0.82	0.07	0.76	0.05	0.30	0.21
Proportion primary industries	0.39	0.16	0.44	0.15	0.51	0.16
Proportion manufacturing	0.11	0.06	0.13	0.06	0.08	0.05
Proportion Catholic	0.17	0.06	0.22	0.06	0.23	0.05
Proportion literate	0.90	0.04	0.96	0.00	0.88	0.14
Proportion border areas	0.00	0.00	0.26	0.44	0.16	0.37
Average distance to capital (km)	130	79	184	88	675	619
Number of districts	30		35		37	

Source: 1901 Colonial Censuses, as available from the Australian Data Archive.

supportive were King with 34 per cent and Gowen with 36 per cent. Parry increased to 36 per cent and Cook increased to 41 per cent. In 1899 there were fewer districts with 100 per cent support—only Thoulcanna and Yantara, but again those districts had few votes. There were nine districts with support for federating over 95 per cent, including Perry, Wakool, Cadell, Nicholson and Boyd all of which had substantial populations. There are 35 districts where support for federating decreased, it was the same in one, and it increased in the remaining 100. There were increases of more than 25 per cent in Culgoa, Yungnulgra, Mootwingee, and Yancowinna. The number of votes supportive of federating increased in 126 districts, and the number of votes against federating increased in 93 districts. Net support (the difference of yes and no) increased in 110 districts. Similarly, turnout increased in almost all districts and it more than doubled in 14 districts.

The means for the key district-level variables in 1901 are reported by colony in Table

3.3. The proportion of males, which is always more than half, is highest in Western Australia. The proportion of adults aged 21–29 is fairly similar across the colonies, however slightly higher in Western Australia. There is more variation across states in the share of native-born (those born within the colony), with a particularly low share in Western Australia. New South Wales is the colony with the highest share in primary industries, which includes agriculture and mining, with manufacturing being especially high in Victoria and South Australia. The proportion Catholic varies a little around the average, being lowest in South Australia, followed by Tasmania. Among adults, the proportion literate (able to read and write) is around 90 per cent; higher in Victoria and lower in Queensland and Western Australia. Around a quarter of the districts border with other states, except for Tasmania which is an island. And with the exception of South Australia, the average distance from the state capitals largely reflect the differences in overall size.[26] Our analysis is conducted on a district basis, and so the number of districts determines the number of observations in any particular year-colony combination.

3.5 Analysis

3.5.1 Model

We are interested in examining the demographic, economic, and geographic factors that were associated with support for federation. Our regression model is:

$$\begin{aligned}\text{support} =& \text{state} + \text{year} + \text{male} + \text{native} + \text{young} + \text{primary industry} + \text{manufacturing} + \\ & \text{catholic} + \text{literacy} + \text{turnout} + \text{border} + \ln(\text{distance}).\end{aligned}$$

Where support is the proportion of voters in a district that were supportive of federating; state is a colony-specific fixed effect with possible values: QLD, NSW, VIC, TAS, SA, and WA; year is year-specific fixed effect with possible values 1898, 1899, and 1900; male

[26] In South Australia electoral districts are used and these reflect population size. As a consequence, the vast and largely empty regions of the northern territory and the north and west of the southern territory carry little weight. Thus most of the districts are clustered in the south eastern corner, with seven of the 27 districts less than 10km from the centre of Adelaide.

is the proportion of males in the district; native is the proportion of those that were born in that colony; young is the proportion of voting age adults in the district that are aged 21 to 29; primary industry is the proportion of those in the district that are employed in primary industries; manufacturing is the proportion of those in the district that are employed in manufacturing; catholic is the proportion of those in the district that are catholic; literacy is the proportion of those in the district that are able to read and write; turnout is the proportion of eligible voters in that district who actually voted; border is a dummy variable that is one if the district is on a border and zero otherwise; and finally, distance is a the number of kilometres between the district and the capital of the colony. There are various aspect to be aware of including: that only Western Australia voted in 1900 and that the vote was after Federation had been agreed to; and that the number of kilometres for the district that is the capital of the colony itself was set at one.

3.5.2 Regression results and discussion

Table 3.4 reports the results of OLS estimation of the proportion of yes votes in a district. It is important to caution that these are associations and not necessarily causal effects. Our intention is to explore the strength and size of the associations over range of different variables rather than to try and estimate a causal relationship for one specific variable or channel of influence. As the dependent variable is a proportion, strictly speaking, it would be more appropriate to use beta regression, however these results can be more difficult to interpret. The results of beta regression are provided in Table 3.5, however the significance and signs of the coefficients are essentially unchanged and the magnitude of the effects are similar. To aid interpretability, the coefficients in Table 3.5 have been converted to marginal effects

The baseline results are estimated with all colonies and years, using fixed effects for colony and year as needed. There are no within-district dependent or explanatory variables, and the colony-level fixed effects should go some way to accounting for correlations within each colony, as such, following Abadie et al. (2017) our standard errors are unclustered. The other regressions examine specific colonies, or groups of colonies, to examine

Table 3.4: Results using OLS

	Dependent variable: Proportion support for Federation				
	All	QLD only	NSW only	VIC only	TAS/SA/WA
QLD	−0.142***				
	(0.021)				
SA	0.260***				
	(0.024)				
TAS	0.406***				0.182***
	(0.024)				(0.032)
VIC	0.335***				
	(0.019)				
WA	−0.190***				−0.393***
	(0.034)				(0.063)
Is 1899	0.072***		0.028	0.083***	0.131***
	(0.013)		(0.018)	(0.017)	(0.023)
Male	−0.481***	0.421	−0.536***	−0.610**	−0.536**
	(0.131)	(0.661)	(0.175)	(0.290)	(0.270)
Aged 21 to 29	−0.169	−0.203	−1.036***	0.169	0.193
	(0.144)	(0.409)	(0.252)	(0.315)	(0.255)
Born in colony	−0.747***	−0.484	−0.489***	−0.565***	−0.954***
	(0.055)	(0.409)	(0.078)	(0.156)	(0.122)
Primary industries	0.055	−0.357	0.101	−0.075	0.189
	(0.068)	(0.265)	(0.101)	(0.061)	(0.117)
Manufacturing	−0.065	−0.657	−0.514		0.210
	(0.197)	(0.748)	(0.331)		(0.311)
Catholic	0.084	−0.176	0.089	−0.029	−0.283
	(0.081)	(0.339)	(0.109)	(0.122)	(0.178)
Literacy	0.076	−0.019	0.826***		0.270*
	(0.088)	(0.261)	(0.200)		(0.161)
Turnout	0.081	0.223	0.230**	0.197	−0.123
	(0.062)	(0.300)	(0.089)	(0.133)	(0.103)
Is border	0.006	0.025	0.065***	−0.022	0.025
	(0.014)	(0.052)	(0.019)	(0.018)	(0.041)
Distance	0.063***	0.089***	0.062***	0.043***	0.050***
	(0.007)	(0.023)	(0.016)	(0.012)	(0.011)
Constant	0.931***	0.347	0.318	1.287***	1.164***
	(0.156)	(0.564)	(0.275)	(0.249)	(0.312)
Observations	556	63	272	70	151
R^2	0.656	0.761	0.613	0.664	0.684
Adjusted R^2	0.646	0.715	0.597	0.613	0.654
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 3.5: Results using beta regression

	Dependent variable: Proportion support for Federation				
	All	QLD only	NSW only	VIC only	TAS/SA/WA
QLD	−0.172***				
	(0.103)				
SA	0.198***				
	(0.120)				
TAS	0.293***				
	(0.125)				
VIC	0.257***				
	(0.106)				
WA	−0.301***				−0.515***
	(0.157)				(0.369)
Is 1899	0.068***		0.024	0.079***	0.094***
	(0.069)		(0.092)	(0.135)	(0.139)
Male	−0.52***	0.669	−0.673***	−0.633***	−0.203
	(0.726)	(2.878)	(0.960)	(2.371)	(1.604)
Aged 21 to 29	−0.048	−0.281	−0.835***	0.257	−0.071
	(0.789)	(1.956)	(1.403)	(2.447)	(1.666)
Born in colony	−0.80***	−0.479	−0.646***	−0.64***	−0.787***
	(0.307)	(1.797)	(0.437)	(1.332)	(0.755)
Primary industries	0.112***	−0.371*	0.167*	−0.059	−0.077
	(0.234)	(1.148)	(0.509)	(0.470)	(0.604)
Manufacturing	0.092*	−0.610	−0.491		−0.426*
	(0.310)	(3.125)	(1.720)		(1.561)
Catholic	0.023	−0.118	0.070	−0.075	0.004
	(0.387)	(1.400)	(0.519)	(0.957)	(0.975)
Literacy	0.052	−0.209	0.772***		0.080
	(0.483)	(1.215)	(1.014)		(1.039)
Turnout	0.044	0.339	0.173*	0.203**	−0.162*
	(0.327)	(1.400)	(0.469)	(1.045)	(0.608)
Is border	0.010	0.041	0.059***	−0.026**	−0.089***
	(0.075)	(0.233)	(0.101)	(0.136)	(0.216)
Distance	0.002***	0.001***	0.000***	0.001***	0.005***
	(0.033)	(0.094)	(0.075)	(0.093)	(0.059)
Constant	2.758***	−0.338	0.323	7.048***	5.257***
	(0.827)	(2.377)	(1.414)	(1.964)	(1.969)
Observations	556	63	272	70	151
R^2	0.665	0.779	0.601	0.766	0.562

Note: *p<0.1; **p<0.05; ***p<0.01

whether these baseline results are being driven by specific colonies. Tasmania, South Australia and Western Australia are grouped together for this regression because of the small number of districts in each.

The baseline results are estimated relative to NSW in 1898, so we find that Queensland and Western Australia were less supportive than New South Wales, while the other colonies—South Australia, Tasmania, and Victoria—were more supportive. The coefficient for 1899 is positive, reflecting the increased support for federating in the later vote. There is no separate value for 1900 as Western Australia was the only colony to vote in that year. The second round of referendums is perhaps less interesting because the results for the previous round were known and the likely outcomes in the second round were more of a foregone conclusion. But it is worth seeing if the associations were similar between the first and second rounds and also whether those for Queensland and Western Australia, which did not vote in the first round, resembled those of the other colonies. It is also of value because the 1901 censuses were conducted closer to the 1899 and 1900 votes. As these results are similar, it may provide some reassurance around the reasonableness of using explanatory variables from 1901 for 1898 as well. The main difference is a general upward shift in the proportion of support for federating.

An increasing proportion of males in a district is negatively correlated with support for Federation, which is a surprising result given that women were not uniformly able to vote. This suggests that women may have exerted a positive influence on support for federation, even where they did not have the vote. This may be related to the push for women's suffrage in Australia, which was active at this time. Each colony extended the vote to women at a different time and the Commonwealth granted voting rights to women at a federal level in 1902, which meant that the Commonwealth actually extended the right ahead of several colonies. Specifically, women received the right to vote in South Australia and Western Australia in 1895 and 1899, respectively; New South Wales and Tasmania quickly followed Federation, in 1902 and 1903, respectively; and Queensland and Victoria were the final states at 1905 and 1908, respectively. Sawer (2001) describes how 'South Australian delegates to the Constitutional Convention insisted on a clause in the

Constitution (s41) that would prevent South Australian women voters being denied the Commonwealth franchise'. Sawer argues this made it 'inevitable' that Federation would lead to women's suffrage Australia-wide; hence to a certain extent, supporting federating was supporting women's suffrage. However this coefficient is not uniformly negative, for instance we find that it is positive for the Queensland-only regression, and we do not know the extent to which this factor was front of voters' mind. Additionally, although a petition was lodged (Womanhood Suffrage, 1897), it is difficult to find evidence of how concerted a campaign there was at the time associating women's suffrage with federating.

It may be expected that younger voters would be more likely to be supportive of federating, however we find a negative coefficient in many cases. The exceptions are Victoria and the group of three colonies.

The proportion of adults that were born in the colony is negatively correlated with support for federating. This may be surprising and suggests that support for federating was stronger among those who were either born in another of the Australian colonies or overseas. It may be that those voters were more geographically mobile, had less sense of 'colony-identity', or that those from outside the colony were less committed to colonial autonomy and more willing to see the benefits of integration. In any event, legislation to restrict non-European migration passed the federal parliament in its first year of existence, although we are unable to tell the extent to which this was a decisive issue at the time.

The key variables representing the sectoral interests is the share of the labour force in manufacturing and the share in primary industries. If federating implied greater protection for manufacturing, then the coefficient should be positive, but it is not. The negative coefficient indicates that districts that were relatively industrial did not favour federation. As noted earlier, this could be because even though manufacturers expected a lower post-federation tariff, labour organisations feared that merging with other colonies could undermine union bargaining strength. It may also simply be that the sector was fairly small in comparison to the total labour force. In the main regression, the coefficient on primary industries is positive, but neither variable is significant. Even independent of tariffs, areas that were more reliant on primary industries would have been expected

to be more supportive of federating for the possibility of better infrastructure and water management. It may be that this variable is not disaggregated enough to appropriately capture the difference between, say, mining and farming. Similarly, it may have been that given the different tariffs imposed by each colony that the baseline estimation averages out an effect, but there is no effect in either of the regressions that are specific to Victoria or New South Wales. Tariffs may be either trade diverting or trade creating, however, as discussed by Irwin (2006), essentially the only barrier to inter-colony trade was the tariff; and it may be that the free flow of capital and labour was sufficient so that the changed tariffs were not expected to have a significant effect.

The coefficient on the proportion of Catholics is small and insignificant but the significant positive coefficient on literacy is consistent with the idea that those with more education and better information were more likely to favour federating. The coefficient on turnout is generally positive, however is negative in the case of the regression focused on the three colonies grouped together.

Finally, it seems that voting for federating also had an important geographic component. The dummy for the districts of New South Wales that bordered on other colonies takes a positive coefficient that is significant at the 1 per cent level. So long as the census variables were reasonably constant this would be consistent with the view that a lack of transport integration and other administrative barriers were a costly irritation that federating could overcome (Pringle, 1978, p. 235). By contrast the border dummy takes a negative coefficient in Victoria, which is opposite to that of New South Wales. This hardly seems consistent with the prospect of gains from interaction across the borders but could be consistent with the implied reduction of the reduction in tariff protection for Victoria at the border with New South Wales. For the remaining colonies the border dummy is insignificant. This is not surprising as most of the border districts are in relatively arid areas and much of the commerce with other colonies was conducted by coastal shipping.

The influence of geographical location was not simply an issue at the border. The coefficient on the log of distance from the respective colonial capital city is positive and significant, suggesting that support for federating increased with remoteness from the seat

of colonial power. Of particular note is the strongly significant positive coefficient on log distance from the colonial capital Brisbane. While this has sometimes been associated with tensions between the mining districts (particularly gold mining) and the populations further south, the coefficient is larger than those for the other colonies, where discord between mining and other interests was much less marked. Although the effect of distance is remarkably strong, this has received little attention in the literature on the referendums and underlying reasons for it remain unclear.

3.6 Conclusion

Our analysis of spatial voting patterns in the referendums that lead to Federation yields four key findings. The first is that, under the assumption that the uniform tariff would look more like that in Victoria than in New South Wales, the association between votes and the sectoral composition seems not to support the predictions of customs union theory. Second, there is a strong correlation between the proportion of migrants in a district and support for federating, something that has not been emphasised in most of the literature. Thus support for federating seems to be related more to migration than to trade. Third, distance from the metropolitan centre of each colony is strongly correlated with support for federating, a finding has also been overlooked in the literature and is open to interpretation. Finally, for New South Wales, support for federating is positively correlated with the proportion of females in the district and the proportion literate.

It is worth emphasising two final points in relation to the historical literature on the referendums. The first is that those studies that have examined the distribution of support for federating have focused on the varying effects of economic structure, politics and persuasion. But the key associations identified here, the shares of migrants and females, as well as distance from the capital have received little attention. It would be useful to gain a deeper understanding of these apparently strong relationships. Second, some of the more recent contributions raise the question of the timing of Federation, specifically why it occurred at the end of the 1890s rather than a decade or two earlier (or later). Our results suggest that perhaps societal changes in attitudes and aspirations

were more important than differences in economic or political structure.

Future work should look further into the relationship between support for federating and the women's suffrage movement, especially in New South Wales. Future work could also make more of a difference between migrants born in an Australian colony and those born overseas, as at the moment they are both treated as having not been born in the colony. Finally, if it were possible to overlay banking or incomes data then a variety of additional questions could be examined.

Chapter 4

A Surname-Based Analysis of Tasmanian Social Mobility

4.1 Introduction

Surname analysis suggests a low level of social mobility in Tasmania, Australia. By way of background, social mobility considers how a person's social status is related to the status of their ancestors. If one's ancestors play a large role in determining one's outcomes, then the level of mobility may be low. This would have implications for the type of economic policies that are appropriate. For instance, to counteract this public education funding may need to be higher, and tax and transfer settings may need to be more progressive. It also has implications for the type of economic policies that are implemented, because it may be that policies are made by those with entrenched interests in maintaining existing power structures. In terms of time since European-settlement Australia is younger than many of the other countries whose social mobility has been analysed using surnames, and in some ways, nineteenth century Tasmania could be considered a frontier economy. As such, it is of interest whether the low levels of social mobility observed in older, more established, economies is also found in Tasmania.

The nineteenth century part of this paper focuses on individuals born in Tasmania between 1820 and 1899, as before 1820 births were sporadic and relatively rare in Tasmania.

Social status can be assigned to these individuals using data sources such as parliamentary service or attendance at an exclusive school. By comparing the surnames of high-status groups then, with the surnames of high-status groups in the late twentieth century, an estimate can be made of the degree of social mobility in Tasmania.

Our main finding is that over a four-generation period high-status in one generation had a persistence of 0.8 with high-status in the next generation. We examine the meaning of this result to find that it is not unexpected although it is higher than estimates of intergenerational elasticity in Australia, such as Deutscher and Mazumder (2019). That is, our finding, is that these days the proportion of certain surnames in high-status groups is fairly similar to the proportion in the nineteenth century. This is similar to the results of Clark (2014), who find similar results focusing mostly on the U.S. and the U.K. but also consider other countries. Our results are much the same as the estimates of Clark et al. (2017) who applied this method in Australia with datasets focused on university education and doctors.

4.2 Pre-Federation Tasmania

Social mobility is especially important in Tasmania because of its history as a penal colony. Tasmania is an island state of Australia that was inhabited by Indigenous peoples at least 35,000 years before European settlement. The Dutch were the first known Europeans to name the island, and by 1777, when it was sighted by the British Captain James Cook, it was called Van Diemen's Land. It is estimated that before British settlement the Indigenous population of Tasmania was between 5,000 and 10,000 persons (ABS, 1996). In terms of land area, Tasmania is of similar size to West Virginia in the USA, or Ireland in Europe.

British colonisation began in the south-east of Tasmania in 1803 with a settlement at Risdon Cove, followed in 1804 by Sullivans Cove, around which the present-day capital, Hobart, formed (ABS, 1996). Most of the initial European population of Tasmania were either convicts or guards, and their settlement decimated the Indigenous Tasmanian population. Few Indigenous peoples survived beyond 1830, and none by 1876 (ABS, 2002).

The Gold Rush in Victoria, beginning in 1851, saw many people leave Tasmania. Convict transportation to Tasmania ceased soon after, in 1853, possibly due to a change of government in Britain (McLean, 2013), and in 1856, the name of the colony was changed to Tasmania when it became self-governing. On 1 January 1901 Tasmania joined with five other colonies to form Australia.

Records from the 1842 Tasmanian Census suggest that of the 57,420 persons recorded as being in the colony, only 27,216 persons are recorded as being 'Born in the Colony' or 'Arrived Free'.[1] Convicts not only provided labour, but were also consumers (Meredith and Oxley, 2014, p. 113). Moyle (2015) describes the nineteenth century Tasmanian economy as 'predominantly agricultural' in its early days, while by 'the end of the 1860s... wool accounted for around half of Tasmania's export income'

4.2.1 Population data

A dataset of Tasmanian births for the nineteenth century is available from LINC, a collaboration of Tasmanian libraries and archives, via `data.gov.au`. Aspects of this dataset have been used as a component of the 'Founders and Survivors Project' that traces the histories of individuals and families in Australia (Bradley et al., 2010). An example of a birth entry is shown in Figure 4.1. The first entry is for a child born on 19 November 1834, named Christiana Susanna.

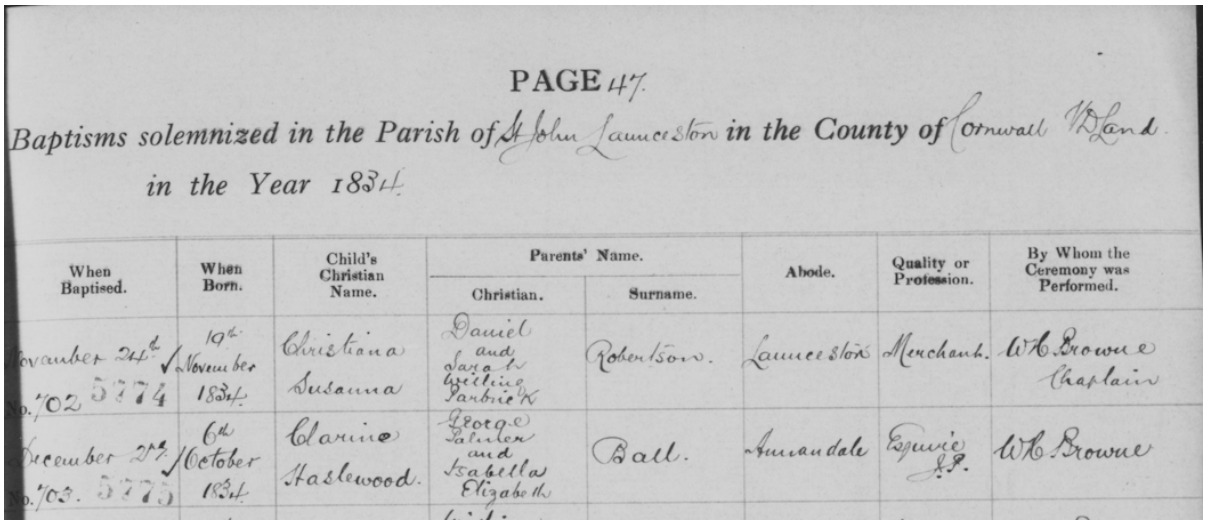
There are 211,604 entries in the dataset recorded as being born between 1820 and 1899 (Figure 4.2).[2] Figure 4.2 shows the substantial impact of the Gold Rush in Victoria, which began in 1851. The decline beginning in the late 1890s may be due to responsibility for the birth registers being taken over by the Commonwealth.

Other authors, such as Moyle (2015) have used datasets of Tasmanian births to analyse Tasmanian fertility rates for this period. For instance, Moyle (2015) finds that fertility began declining in the late 1880s. This paper estimates social mobility and following the

[1] To reconstruct this number, obtain the data from 'Historical Census and Colonial Data Archive' at `http://hccda.ada.edu.au/pages/TAS-1842-census-01_1`. In the 'Civil Condition' columns sum 'Males' and 'Females' for both 'Born in the Colony' and 'Arrived Free'.

[2] There are 2,127 birth records for years before 1820, however the population was small at this time and there are few of the other records that would be needed to analyse social mobility for these births.

Figure 4.1: Example of church baptism record

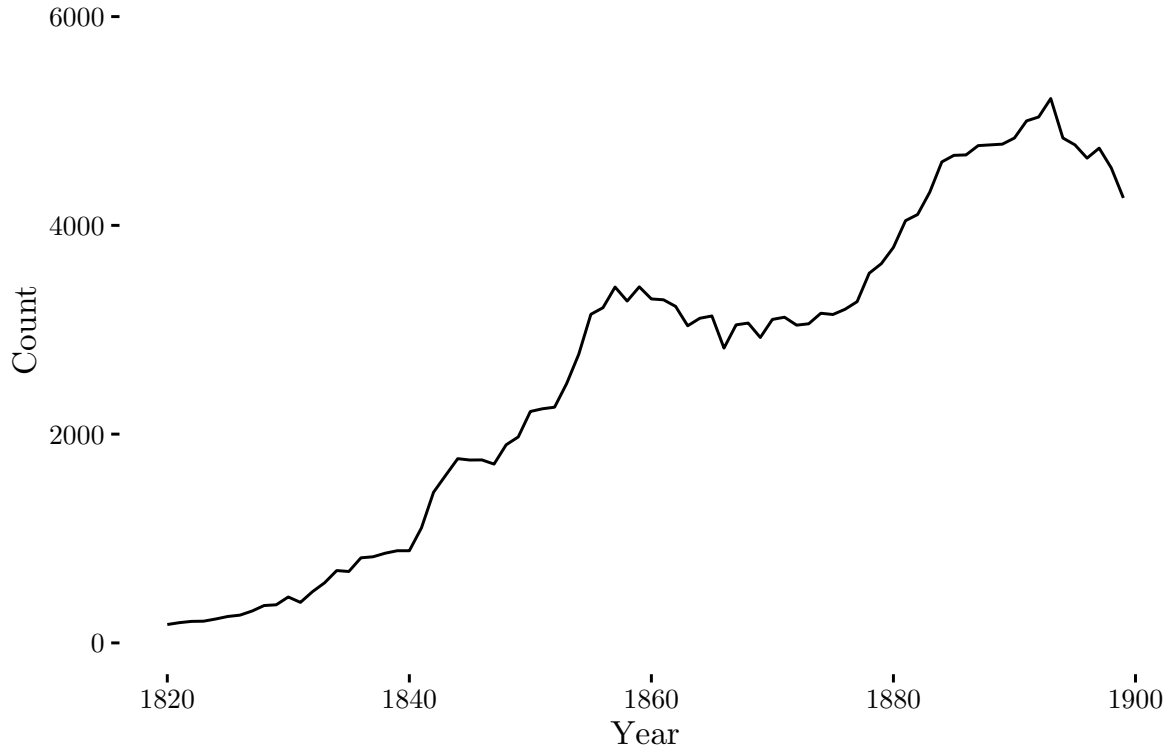


PAGE 47.

Baptisms solemnized in the Parish of St John Launceston in the County of Cornwall Land in the Year 1834

When Baptised.	When Born.	Child's Christian Name.	Parents' Name. Christian.	Parents' Name. Surname.	Abode.	Quality or Profession.	By Whom the Ceremony was Performed.
November 24th	19th November	Christiana	Daniel and Sarah Isabella Patrick	Robertson.	Launceston	Merchant.	W. H. Browne Chaplain
No. 702 5774	1834.	Susanna					
December 21st	6th October	Clarina	George Palmer and Isabella Elizabeth	Ball.	Launceston	Esquire J.P.	W. H. Browne
No. 703. 5775	1834	Haslewood.					

Figure 4.2: Annual birth counts, 1820 – 1899



method of Clark (2014) focuses on the surnames in the dataset. The 35-most-common surnames based on births data between 1820 and 1899 are summarised in Table 4.1.

Smith is the most common surname in the dataset over the 1820 to 1899 period, accounting for around 2 per cent of all births. The next most popular surnames are Williams, Jones, Brown, and Wilson. There are 14,645 different surnames over this period, of which 5,974 appear only once and 9,781 appear five or fewer times. The distribution of the surnames appears similar to a Pareto distribution. Figures 4.3a and 4.3b show the rank of the surname in terms of commonality on the horizontal axis and the number of births with that surname on the vertical axis. Figure 4.3a includes all names, while Figure 4.3b only includes the 100-most-common, less Smith. The frequency of such surnames may be artificially decreased by misspellings and transcription errors. For instance, both Whitefield and Whitefeld appear twice, although it is not clear whether this is accurate.

The most common surnames are reasonably stable over the period 1820 to 1899. One way to see this is to examine the rank of the overall 20-most-common surnames in four twenty-year periods: 1820–1839, 1840–1859, 1860–1879, 1880–1899 (Figure 4.4). Note that these groupings are just for illustrative purposes. If a surname is equally common in each twenty-year period, as measured by ranking, then the line corresponding to that surname in Figure 4.4 will be horizontal. It can be seen that most of the lines are either entirely, or close to, horizontal.

One concern with using this dataset is the extent of the share of Tasmanian births that it does not contain. Nineteenth century civil registration datasets are rarely complete, and Tasmania is unlikely to be an exception, especially in the first half of the century. As in the UK, Tasmania's early civil registration data were recorded by clergy and kept in parish registers. This made them susceptible to the idiosyncrasies of the clergy, as well as local events such as loss, fire, and flood.

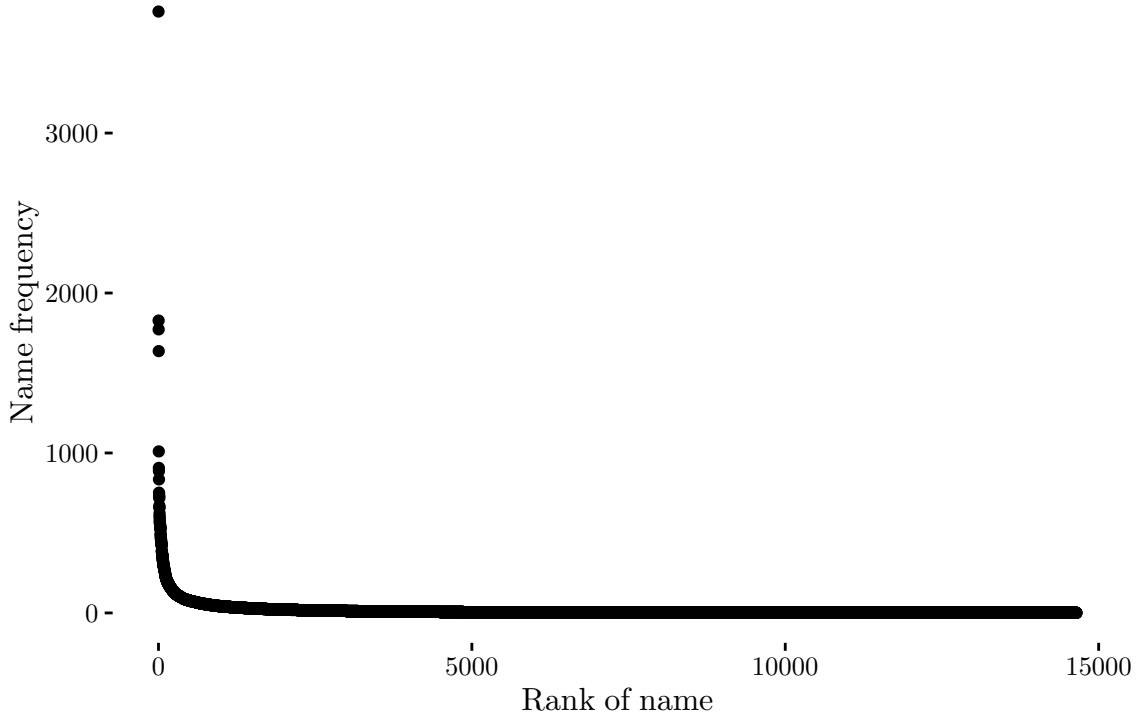
An act of the British Parliament in 1836 established a central office to keep copies of the parish registers in the UK. Kippen (2002a) analyses civil registration in nineteenth century Tasmania, and describes how, following this, in 1838 Tasmania became the first Australian colony to similarly establish a central office. Tasmania was divided into seven

Table 4.1: 35-most-common birth and death surnames, 1820 – 1899

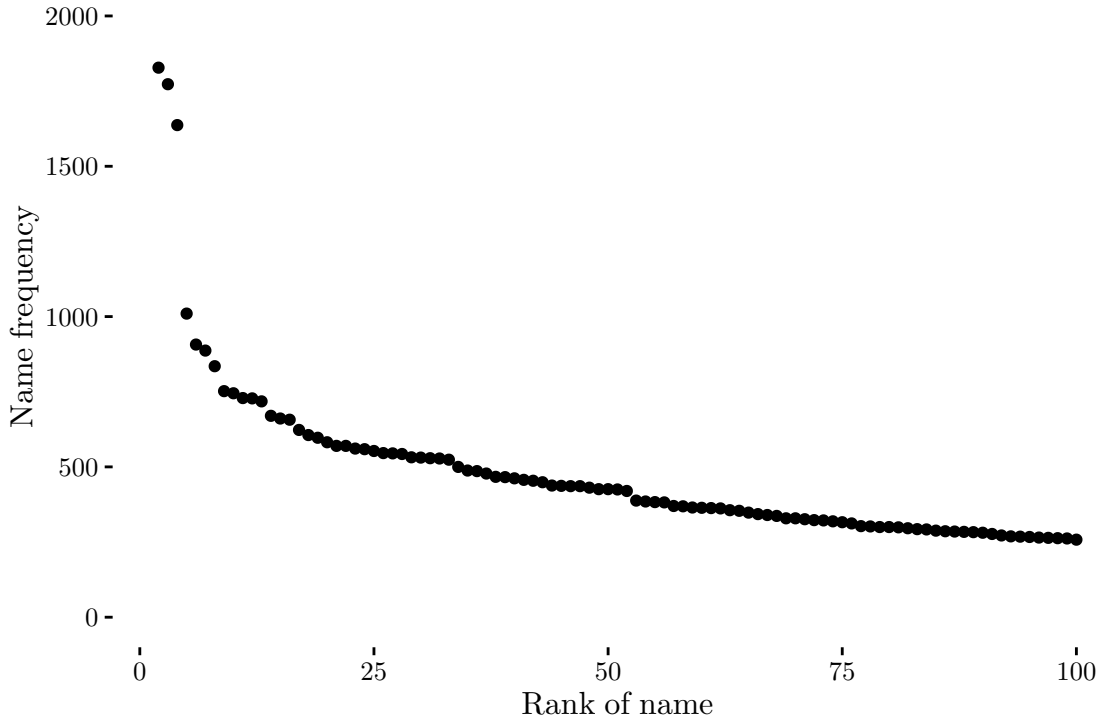
Births			Deaths		
Surname	Frequency	Percentage	Surname	Frequency	Percentage
Smith	3,760	1.8	Smith	1,698	1.8
Williams	1,828	0.87	Williams	841	0.89
Jones	1,773	0.85	Jones	803	0.85
Brown	1,637	0.78	Brown	778	0.83
Wilson	1,010	0.48	Wilson	503	0.53
Taylor	907	0.43	Taylor	492	0.52
Johnson	887	0.42	Johnson	428	0.45
White	835	0.4	White	386	0.41
Davis	752	0.36	Davis	376	0.4
Walker	745	0.36	Thompson	362	0.38
Clark	729	0.35	Wright	310	0.33
Moore	728	0.35	King	300	0.32
Turner	718	0.34	Martin	300	0.32
Harris	670	0.32	Turner	299	0.32
King	661	0.32	Harris	294	0.31
Wright	657	0.31	Thomas	285	0.3
Hall	623	0.3	Hill	282	0.3
Scott	606	0.29	Miller	282	0.3
Hill	597	0.29	Moore	282	0.3
Martin	582	0.28	Green	279	0.3
Edwards	570	0.27	Clarke	276	0.29
Miller	570	0.27	Walker	276	0.29
Evans	561	0.27	Lewis	260	0.28
Thompson	559	0.27	**Robinson**	258	0.27
Young	553	0.26	Clark	254	0.27
Green	546	0.26	Evans	248	0.26
Anderson	545	0.26	Anderson	245	0.26
Lewis	543	0.26	Edwards	242	0.26
Johnston	532	0.25	**Roberts**	242	0.26
Thomas	531	0.25	Watson	241	0.26
Cox	529	0.25	Hall	240	0.25
Clarke	528	0.25	Young	239	0.25
Cooper	524	0.25	**Murphy**	236	0.25
Watson	500	0.24	**Jackson**	221	0.23
Collins	488	0.23	**Kelly**	218	0.23

Figure 4.3: Tasmanian births surname distributions

(a) All surnames



(b) 100-most-common surnames, less Smith



registration districts, each run by a Deputy Registrar who was responsible for recording the details of births, deaths, and marriages amongst the free population. Convict records were kept in separate registers.

Parents were responsible for registering newborn babies, and householders were responsible for registering deaths in their homes. The clergy were required to notify the Deputy Registrar of any burials that occurred without registration. The maximum penalty for non-compliance was £10, at a time when the average annual wage, with board and lodging, for a milkman was £20 and for a stone-cutter was £40 (Emigration from the United Kingdom, 1838, p. 162).

Kippen (2002a) describes compliance with the Act as 'an ongoing problem', and financial incentives were established in 1843 that encouraged Deputy Registrars to accurately carry-out their work. As at 1853, the Registrar 'believed (the) death registration was complete', however births were considered under-reported, for instance in '1847, there were 2,041 baptisms, and only 1,531 registered births' (Kippen, 2002a, p. 49)

One way to help allay concerns around bias in the dataset is to compare it with other records, such as censuses. There are records for eight nineteenth century censuses available at the Historical Census and Colonial Data Archive (HCCDA).[3] These are for the years 1842, 1848, 1851, 1857, 1861, 1870, 1881, and 1891. Although not every census specifies the number of births in the preceding year, estimates can be made (Table 4.2).[4]

The 1861 and 1870 censuses do not provide the 'Born in the Colony' number that is in the four earlier censuses. However, they do contain the number of children less than one-year-old, by sex. This suggests there were 3,117 births in the year to the date of the census in 1861, and 2,887 births in the year to the date of the census in 1870.[5] As with

[3]The HCCDA is a sub-archive of the Australian Data Archive and is available at `http://hccda.ada.edu.au/`.

[4]The first four censuses (1842, 1848, 1851, 1857) describe the number of people 'Born in the Colony', by sex. Comparing this number between the four censuses provides an estimate of the number of births each year. It will be an underestimate because it will be net of those who died, left Tasmania, or were otherwise not included in the census, and will also suffer from being a linear interpolation. Nonetheless, the censuses imply 945 births annually between 1842 and 1848; 1,066 births annually between 1848 and 1851; and 1,428 births annually between 1851 and 1857. By way of comparison, the dataset used in this paper contains 1,443 births in 1842, 1,713 births in 1847, 2,243 births in 1851 and 3,409 births in 1857.

[5]To reconstruct these numbers go to `http://hccda.ada.edu.au/pages/TAS-1861-census-01_1` or `http://hccda.ada.edu.au/pages/TAS-1870-census-01_1`, then the 'Totals' row, and sum the 'Under six months' and 'Under one year' columns (for the 1870 Census only the 'Under One Year' number is

the earlier censuses, these figures are net of deaths and departures and so will be likely an underestimate of the number of births. Nonetheless, the dataset used in this paper has 3,287 recorded births in 1861 and 3,099 in 1870.

There is more detail available for the 1881 and 1891 censuses. In the Introductory Report to the 1881 Census the Superintendent of the Census writes '[b]etween the Census of 1870 and that of 1881, 36,126 births were registered'. Although it is not possible to exactly match the births in the dataset with the date of the census within a year, adding half of the births in 1870 and 1881 as well as those in the intervening years provides a count of 43,670. And in the Introductory Report of the 1891 Census, there are 4,588 infants under one year, which is similar to the 5,001 found in the dataset used in this paper.

Another way to help lessen concerns around bias in the dataset is to compare births with another measure. For instance, a dataset of deaths is available from the same source. The deaths dataset has 94,603 entries in the period 1820 to 1899. The 35-most-common surnames based on deaths data between 1820 and 1899 are summarised in Table 4.1. Almost all of the most-common surnames are present, at similar rates, in both datasets. Surnames that are bold do not appear in both lists. They are concentrated toward the end of Table 4.1 because the overlap is only evaluated on the names in Table 4.1.

The main concern in terms of the completeness of the Tasmanian nineteenth century deaths dataset is infant deaths, as they may not be registered. Although the dataset is subject to this, the impact of it should not be substantial because the estimate of infant mortality in Tasmania at this time is 21–25 deaths per 1,000 live births (de Looper (2014, p. 63) and Kippen (2002b)). Additionally, the sex ratio of births does not seem overly selective (Figure 4.5), despite the considerable skew in the sex ratio of the broader population.

To summarise, the dataset of Tasmanian births contains 211,604 birth records over the period 1820 to 1899. The quality of the dataset, particularly in the second half of the nineteenth century seems likely to be comparable to other records from that time. There will

needed) for both 'Males' and 'Females'.

Figure 4.4: Annual count of available birth records, 1820 – 1899

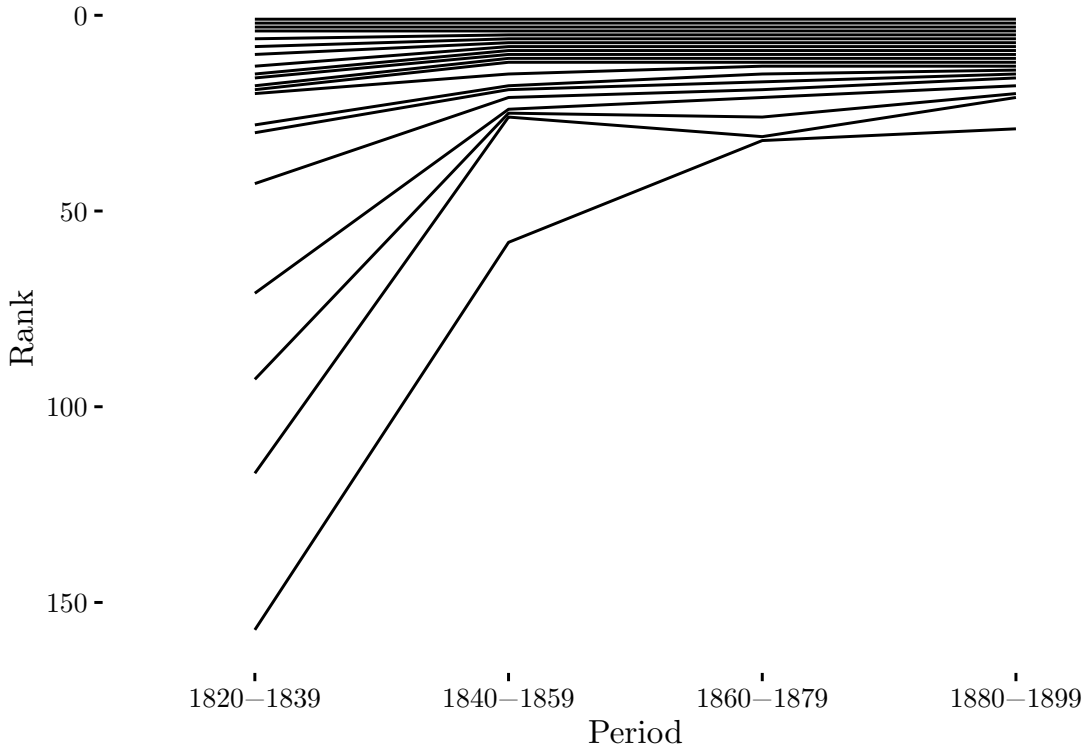


Table 4.2: Annual number of births, based on census

Census	Type	Number	Births dataset
1842 – 1848	Linear interpolation	945	1,443
1848 – 1851	Linear interpolation	1,066	1,713
1851 – 1857	Linear interpolation	1,428	2,243
1861	Census value	3,117	3,287
1870	Census value	2,887	3,099
1881	Linear interpolation	3,612	4,045
1891	Census value	4,588	5,001

be under-recording, particularly due to neo-natal mortality, however the dataset is being used to understand the share of particular surnames in the population and appropriate for this purpose.

4.2.2 Identifying status

A traditional analysis of social mobility would focus on income. For instance, Solon (1992) estimates intergenerational income mobility in the USA using the PSID dataset; Leigh (2007) compares the wages of fathers and sons in Australia using the HILDA dataset; and Deutscher and Mazumder (2019) use Australian tax data. However, as datasets with longer time frames become available there has been more work on multi-generational mobility, summarised by Solon (2015). Often these datasets lack information on incomes or wealth at an individual level and instead analysis focuses on identifiable high-status groups. For instance, Lindahl et al. (2015) use surnames to find more persistence when considering more than two generations than only two generations, in contrast to a 'Buddenbrooks phenomenon'. And analysis of longer time frames using surnames allows examination of the impact of grandparents, for instance, Olivetti et al. (2018). This paper follows those and Clark (2014) in focusing on the surnames of high-status groups.

By examining the surnames that high-status groups are composed of over time, a measure of social mobility can be constructed. For instance, attending an exclusive school or working as a parliamentarian would be considered high-status for the purpose of this paper, given the fees charged by exclusive schools relative to the average income at the time, and the types of people who tended to become parliamentarians.

The key aspect of this analysis, following Clark (2014), is the relative representation of a surname, or group of surnames. That is, the share of that surname/s in the high-status group, compared with the share of that surname/s in the general population. In Tables 4.3 and 4.4, this is the 'Ratio' column. These are not measures of statistical significance, but are instead descriptive and a value greater than one implies that the surname is over-represented in high-status surnames, compared with the population. Almost all the political surnames are over-represented, with the exception of 'Smith' which is underrep-

resented. Similarly, students at the Hutchins School tend to have a different composition to the population.

The composition of a high-status group is subjective as it is meant to include more than just income or wealth, and it is contingent on the data availability. Additional subjectivity is introduced when a decision is made about what constitutes a distinctive surname. For instance, Clark (2014) considers various cut-offs such as fewer than 100 persons holding the surname and also fewer than 500 persons. There is further discussion of this issue later in the paper, but briefly, distinctive surnames are needed to link high-status groups over time, but the decision as to which surnames are distinctive may be influential.

4.2.2.1 Parliamentary service

Tasmania has a bicameral parliament made up of the Legislative Council (Upper House) and the House of Assembly (Lower House). The Upper House is unusual in that members are typically independent, rather than party-affiliated.

Tasmania was initially a territory of New South Wales and only became a separate colony in 1825. After this, what became the Legislative Council met. As described by Korobacz (1971) the members of the Council were typically appointed to one- to three-year terms that were able to be reappointed, from its establishment until 1851. Between 1851 and 1856, two-thirds of the members were elected and one-third were appointed. And from 1856, Tasmania was created as its own colony and all Members of the Legislative Council, as well as the newly established House of Assembly, were elected.

Parliamentary service generally tends to be associated with high-status. For instance, Reynolds (1969, p. 1) argues that '[t]he families that received land grants prior to 1831 continued to play an important role in the economic and political life of the colony until the concluding years of the century'. A record of, and some biographical information about, those who have served in the Tasmanian Parliament is available on a disaggregated basis on the Tasmanian Parliament's website. For the members of parliaments before 1856 the records are from Korobacz (1971) which are made available on the Parliament's

website. For members of parliaments since 1856 the disaggregated information available on the Parliament's website was combined with an aggregated dataset supplied by the Parliamentary Research Service from the Tasmanian Parliamentary Library's Members of Parliament Database.

There are 371 Tasmanian parliamentarians born before 1899 (Figure 4.6). The 35-most-common surnames of politicians born before 1900 are summarised in Table 4.3. Archer, with 12 politicians, is the most common surname, and is much more common in the dataset of politicians, than in the births dataset. The Shoobridge family is another prominent political family. For instance, Louis Manton Shoobridge, who was born in 1851 in Tasmania was the son of Ebenezer, the brother of William, the father of Rupert, and the grandfather of Louis, all of whom were members of a Tasmanian parliament at some stage. The Shoobridge family are additionally related by marriage to at least one other politician, Philip Fysh, although Fysh is not a common surname in its own right. To be clear, two people in the same generation with the same surname who are clearly closely related, for instance brothers, are not treated differently to two people in the same generation with the same surname who are not clearly closely related. So there may be some bias present from changing family size.

4.2.2.2 The Hutchins School

The Hutchins School is a school in Hobart whose first cohort entered in 1846. It is one of the oldest continuously run schools in Australia. Hutchins School student records are available via the 'Roll of Scholars' published in two editions, 1993 and 1996.[6]

From the school records for classes that entered between 1846 and 1899 there are records of 1,558 students. This corresponds to years of birth between 1829 and 1895 (Figure 4.7).

Attendance at the Hutchins School can be considered a signal of being part of a high-status family because of the financial cost and the role played by the parents of the

[6]Digitised records were provided for classes entering up to 1900. The digitisation process is not completed for classes who entered after 1900, and even if it were complete there may be privacy issues that would not make it appropriate to use the dataset.

Figure 4.5: Annual births sex ratio, 1820 – 1899

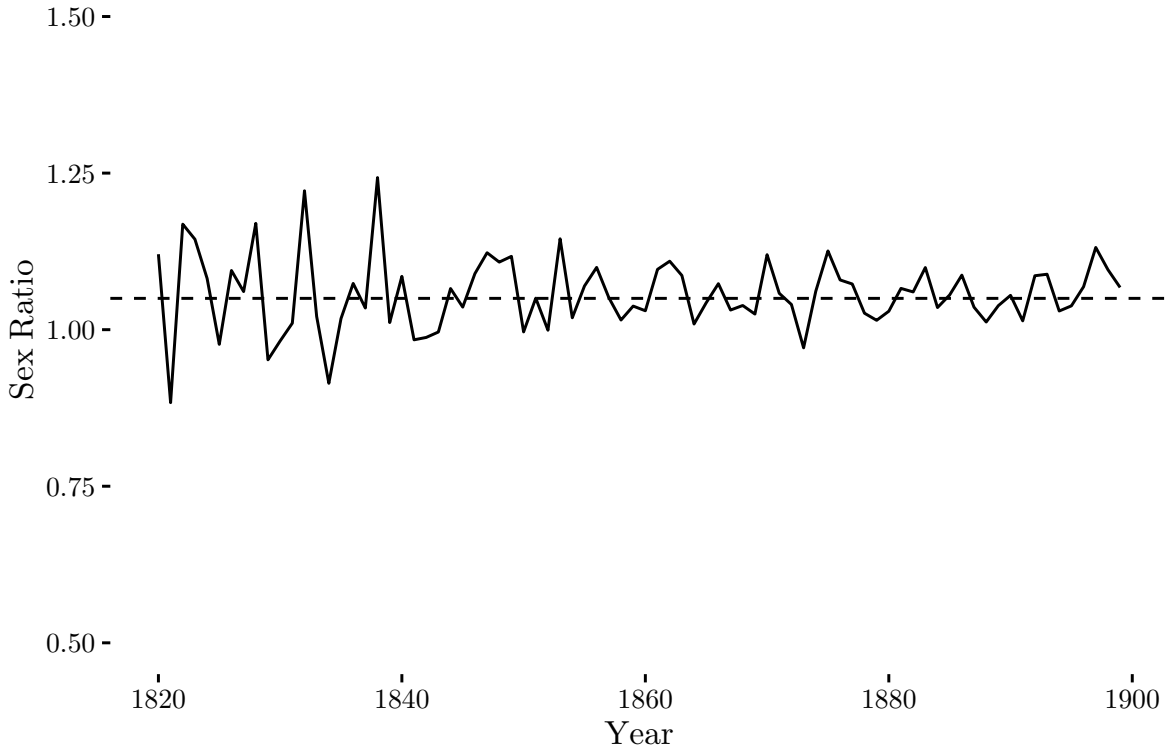


Figure 4.6: Annual count of parliamentarians by birth year

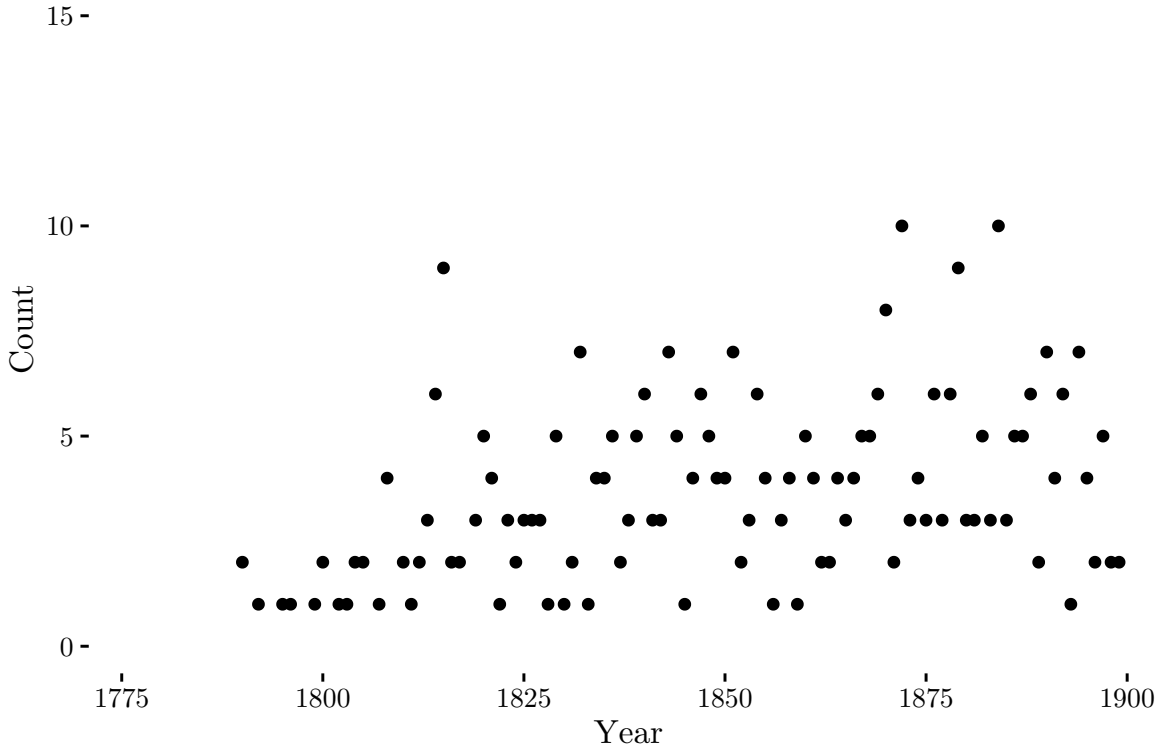
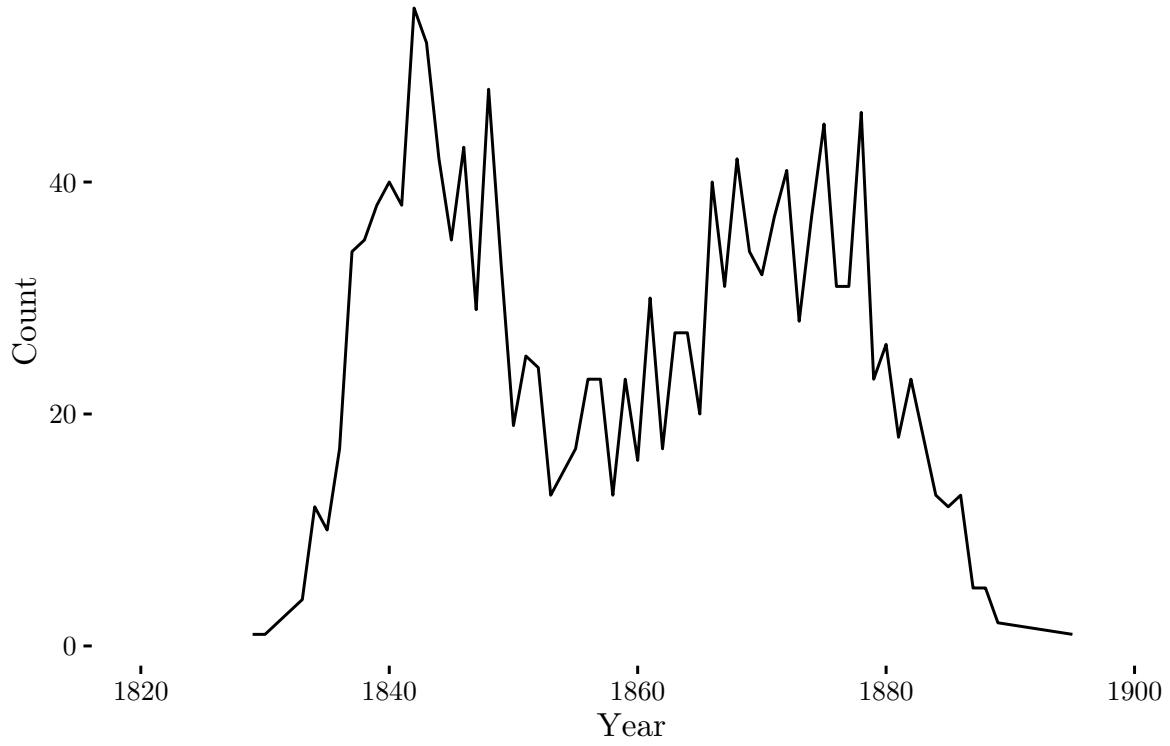


Table 4.3: 35 most-common politician surnames, 1820 – 1899

	Politicians		All births		
Surname	Frequency	Percentage	Frequency	Percentage	Ratio
Archer	12	2.64	329	0.16	16.8
Smith	6	1.32	3,760	1.8	0.74
Chapman	5	1.1	216	0.1	10.66
Gibson	5	1.1	199	0.1	11.58
Murdoch	5	1.1	78	0.04	29.53
Shoobridge	5	1.1	53	0.03	43.46
Brown	4	0.88	1,637	0.78	1.13
Butler	4	0.88	316	0.15	5.83
Cameron	4	0.88	191	0.09	9.65
Gellibrand	4	0.88	23	0.01	80.12
Lord	4	0.88	159	0.08	11.59
Von Stieglitz	4	0.88	2	0	921.38
Bisdee	3	0.66	32	0.02	43.19
Burbury	3	0.66	58	0.03	23.83
Davies	3	0.66	247	0.12	5.6
Dobson	3	0.66	154	0.07	8.97
Douglas	3	0.66	186	0.09	7.43
Fenton	3	0.66	109	0.05	12.68
Field	3	0.66	111	0.05	12.45
Foster	3	0.66	281	0.13	4.92
Giblin	3	0.66	78	0.04	17.72
Grant	3	0.66	194	0.09	7.12
Gunn	3	0.66	76	0.04	18.19
Lewis	3	0.66	543	0.26	2.55
Shaw	3	0.66	256	0.12	5.4
Walker	3	0.66	745	0.36	1.86
Aikenhead	2	0.44	16	0.01	57.59
Anstey	2	0.44	3	0	307.13
Atkins	2	0.44	191	0.09	4.82
Barnes	2	0.44	238	0.11	3.87
Bedford	2	0.44	49	0.02	18.8
Best	2	0.44	204	0.1	4.52
Burgess	2	0.44	337	0.16	2.73
Calvert	2	0.44	70	0.03	13.16
Champ	2	0.44	29	0.01	31.77
Total	122		11,179		

Figure 4.7: Annual count of implied birth year of Hutchins students, 1820 – 1899



Scholars. The fee to attend the Hutchins School in 1846 was £45 School Fees (1846). This was a considerable amount given the average annual wage at the time, discussed earlier, which for a milkman was £20 and for a stone-cutter was £40, and the estimated Tasmania GDP in 1840 was £2,062,000 (Butlin, 1985).

The 35-most-common surnames based on the School Roll for births between 1820 and 1899 are summarised in Table 4.4. As with parliamentarians, many of the 35-most-common surnames of Hutchins students occur in different frequencies in the broader population, as measured by births. For instance, Giblin occurs 13 times in the Hutchins dataset, which is almost 1 per cent of the dataset, while only occurring 78 times in the broader population. Other names that have especially different representations in the Hutchins dataset compared with the births dataset are Westbrook, Perkins, Nicholas, Maxwell, Murdoch, Barclay, Bedford, and Bisdee, all of which are at least ten times as common in the Hutchins dataset than in the broader population.

Table 4.4: 35-most-common Hutchins student surnames, 1820 – 1899

	Hutchins		All births		
Surname	Frequency	Proportion	Frequency	Proportion	Ratio
Fisher	20	0.0130	420	0.0020	6.4758
Smith	19	0.0124	3,760	0.0180	0.6872
Butler	18	0.0117	316	0.0015	7.7463
Jones	14	0.0091	1,773	0.0085	1.0738
Giblin	13	0.0085	78	0.0004	22.6651
Westbrook	13	0.0085	112	0.0005	15.7847
Perkins	12	0.0078	142	0.0007	11.4922
Dobson	10	0.0065	154	0.0007	8.8306
Douglas	10	0.0065	186	0.0009	7.3113
Nicholas	10	0.0065	82	0.0004	16.5843
Evans	9	0.0059	561	0.0027	2.1817
Harris	9	0.0059	670	0.0032	1.8267
Mason	9	0.0059	362	0.0017	3.3810
Maxwell	9	0.0059	63	0.0003	19.4273
Murdoch	9	0.0059	78	0.0004	15.6913
Reid	9	0.0059	383	0.0018	3.1956
Webster	9	0.0059	217	0.0010	5.6402
Wilkinson	9	0.0059	197	0.0009	6.2128
Abbott	8	0.0052	157	0.0008	6.9295
Clarke	8	0.0052	528	0.0025	2.0605
Murphy	8	0.0052	292	0.0014	3.7258
Walker	8	0.0052	745	0.0036	1.4603
Adams	7	0.0046	343	0.0016	2.7753
Barclay	7	0.0046	34	0.0002	27.9981
Bedford	7	0.0046	49	0.0002	19.4273
Bisdee	7	0.0046	32	0.0002	29.7480
Fitzgerald	7	0.0046	122	0.0006	7.8028
Fleming	7	0.0046	167	0.0008	5.7002
Martin	7	0.0046	582	0.0028	1.6356
Reynolds	7	0.0046	277	0.0013	3.4366
Roberts	7	0.0046	478	0.0023	1.9915
Young	7	0.0046	553	0.0026	1.7214
Allen	6	0.0039	382	0.0018	2.1360
Chapman	6	0.0039	216	0.0010	3.7775
Clark	6	0.0039	729	0.0035	1.1193
Total	331		15,240		

4.3 Late twentieth century Tasmania

The second aspect to the method of social mobility estimation used by Clark (2014) is the distribution of surnames in the late twentieth century. If those surnames that were over-represented in nineteenth century high-status groups are still over-represented today, then it may be that social mobility is low.

4.3.1 Population data

Population data for late twentieth century Tasmania are difficult to obtain. For instance, complete datasets of late twentieth century birth or death records do not appear to be available online.

The most recent Tasmanian electoral roll that is available online is the 1980 via Ancestry.com. That electoral roll cannot be downloaded, and can only be queried on a name-by-name basis. The up-to-date electoral roll is only able to be viewed in person at an Australian Electoral Commission office and electronic copying is not allowed. As mentioned by Clark (2014) the Intellectual Property Agency of the Australian Government has made the electoral roll able to be searched, by surname, however the results are for the whole of Australia and cannot be restricted to a particular state. The White Pages, Australia's phone book, is also available to be queried on a name-by-name basis, however they can be restricted to only include Tasmania. Many other countries retain individual census records, and names data can be obtained from that, however in Australia these are destroyed after the responses are obtained and those data are not available.

There are extensive problems with using the electoral roll and the White Pages and neither source is representative of the population. For instance, to be on the electoral roll requires being at least a certain age and having registered. Similarly, the White Pages, requires a fixed address, and a landline.[7] Nonetheless, we use the White Pages here due

[7]The different population measures used for late twentieth century Tasmania, compared with those used for nineteenth century Tasmania, could introduce inaccuracy. For instance, the births dataset is created by aggregating annual flows of surnames. It has not been adjusted for arrivals, deaths and departures, and is necessarily focused on newborn children. Electoral roll or White Pages data, on the other hand, represent counts as at a particular time. Neither the electoral roll or the White Pages should contain many children. Future work should improve on this population measure. As a first step, expanding this work to consider the whole of Australia would at least allow the up-to-date electoral roll

to data availability.

4.3.2 Identifying status

As with the population measure, and with the exception of parliamentarians, the identifiers of status that were used for nineteenth century Tasmania are not available for present-day Tasmania. Examples of signals of status that are publicly available for late twentieth century Tasmania include: service as a parliamentarian, being appointed to the Order of Australia, and working in the legal profession.

4.3.2.1 Order of Australia

Data on those who have been appointed to the Order of Australia are available from its website.[8] The database appears to document almost all recipients of Australian Honours.

The focus of this analysis is on the Order of Australia, which began in 1975. It is divided into the Civil or General Division, and the Military Division, and comprises five levels, in decreasing selectivity: Knight/Dame of the Order (AK/AD); Companion of the Order (AC); Officer of the Order (AO); Member of the Order (AM); and Medal of the Order (OAM).

The Knight/Dame level of the Order has several exceptions. Firstly, it is not applicable in the Military Division. And, secondly, there has only been provision for it in the General Division between 1976 and 1986, as well as between 2014 and 2015. However, there have only been 19 appointments to this level, two of whom are members of the British Royal Family.

As of 2015 there have been 1,308 people appointed to the Order of Australia who listed their location as Tasmania at the time of the appointment (Table 4.5).

As with the births dataset for nineteenth century Tasmania, the most common surname for those living in Tasmania is Smith, with around 2 per cent of the appointments. Other popular surnames are Green, Wilson, Harris, Scott, and Davis (Table 4.6).

look-up tool to be used. While that analysis would not be free of bias, it would be reduced, compared with the level here. Additionally there is a potential for partnerships with the Australian Taxation Office, in a similar manner to Barone and Mocetti (2016) or Deutscher and Mazumder (2019).

[8] https://www.itsanhonour.gov.au/honours/honour_roll

Table 4.5: Order of Australia appointments, by level

Level	Australia-wide	Tasmania only
Dame/Knight of the Order of Australia	19	1
Companion of the Order of Australia	398	9
Officer of the Order of Australia	2,682	69
Member of the Order of Australia	9,488	328
Medal of the Order of Australia	21,746	901

One concern with the Order of Australia dataset is that the lowest level, Medal of the Order of Australia, may not be indicative of status. This level is often awarded for community service, such as running a local sports club. Although this is important, and arguably what the awards should be used to recognise, it is not clear that it is a signal of status. As such, Table 4.6 also shows the surname analysis excluding Medal of the Order of Australia recipients.

4.3.2.2 Legal professionals

Legal professionals are used by Clark (2014) as indicative of status. A similar collection for present-day Tasmania can be constructed because the Law Society of Tasmania publishes a list of its members. The Law Society of Tasmania is a professional association for Tasmanian legal professionals and is part of the Law Council of Australia. Although membership is not compulsory, many Tasmanians connected to the legal profession are members.

A list of current members is available via the Society's website.[9] This indicates there are 622 members of the Society. Analysis of this dataset indicates the most popular surname is Smith, followed by Walker, Brown, and Jones (Table 4.7).

4.3.2.3 Parliament

The dataset of parliamentarians used as an indicator of high-status for the nineteenth century datasets is also available as an indicator of present-day high-status. There are 311 parliamentarians born after 1905. The most popular surnames are Smith, Brown, Hodgman, Archer, and Bacon (Table 4.7). Hodgman, Archer, and Bacon are well-known

[9] https://members.lst.org.au/members/search/people/

Table 4.6: 35 most-common Order of Australia surnames, for Tasmanians

All awardees			Without OAM awardees		
Surname	Frequency	Percentage	Surname	Frequency	Percentage
Smith	24	1.83	Green	5	1.23
Green	10	0.76	Smith	5	1.23
Wilson	9	0.69	Brown	4	0.98
Harris	8	0.61	Walker	4	0.98
Scott	8	0.61	Butler	3	0.74
Davis	7	0.53	James	3	0.74
Brown	6	0.46	Underwood	3	0.74
Burns	6	0.46	Wilkinson	3	0.74
Byrne	6	0.46	Banks	2	0.49
Clark	6	0.46	Barnard	2	0.49
Walker	6	0.46	Benjamin	2	0.49
Bennett	5	0.38	Braithwaite	2	0.49
Butler	5	0.38	Bugg	2	0.49
Davies	5	0.38	Burgess	2	0.49
French	5	0.38	Burns	2	0.49
Osborne	5	0.38	Cameron	2	0.49
Burgess	4	0.31	Canning	2	0.49
Clarke	4	0.31	Colville	2	0.49
Cooper	4	0.31	Cox	2	0.49
Cunningham	4	0.31	Davies	2	0.49
Fisher	4	0.31	Edwards	2	0.49
Foster	4	0.31	Fenton	2	0.49
Jones	4	0.31	Fitzgerald	2	0.49
Kearney	4	0.31	French	2	0.49
King	4	0.31	Gibson	2	0.49
Matthews	4	0.31	Gray	2	0.49
Mitchell	4	0.31	Hughes	2	0.49
Reid	4	0.31	Kearney	2	0.49
Roberts	4	0.31	Knight	2	0.49
Shaw	4	0.31	Melick	2	0.49
Valentine	4	0.31	Miller	2	0.49
Viney	4	0.31	Mitchell	2	0.49
Wilkinson	4	0.31	Morris	2	0.49
Williams	4	0.31	Newell	2	0.49
Banks	3	0.23	Norris	2	0.49
Total	196			84	

Table 4.7: 35 most-common legal profession and political surnames

Legal profession			Politicians		
Surname	Frequency	Proportion	Surname	Frequency	Proportion
Smith	6	0.0096	Smith	5	0.0161
Walker	6	0.0096	Brown	4	0.0129
Brown	5	0.0080	Hodgman	4	0.0129
Jones	4	0.0064	Archer	3	0.0096
Bartlett	3	0.0048	Bacon	3	0.0096
Chan	3	0.0048	Barnard	3	0.0096
Davies	3	0.0048	Batt	3	0.0096
Dixon	3	0.0048	Gibson	3	0.0096
Edwards	3	0.0048	Green	3	0.0096
Green	3	0.0048	Groom	3	0.0096
Groom	3	0.0048	Hiscutt	3	0.0096
Johnson	3	0.0048	Marriott	3	0.0096
Mitchell	3	0.0048	O'Byrne	3	0.0096
Tan	3	0.0048	Armstrong	2	0.0064
Topfer	3	0.0048	Barker	2	0.0064
White	3	0.0048	Barnett	2	0.0064
Williams	3	0.0048	Beattie	2	0.0064
Wood	3	0.0048	Bessell	2	0.0064
Zeeman	3	0.0048	Best	2	0.0064
Ayliffe	2	0.0032	Braid	2	0.0064
Browne	2	0.0032	Butler	2	0.0064
Cooper	2	0.0032	Coates	2	0.0064
Davis	2	0.0032	Cole	2	0.0064
Eddington	2	0.0032	Davies	2	0.0064
Foon	2	0.0032	Davis	2	0.0064
Foster	2	0.0032	Fry	2	0.0064
Grant	2	0.0032	Harriss	2	0.0064
Griffits	2	0.0032	Hope	2	0.0064
Gunadasa	2	0.0032	Jackson	2	0.0064
Gunson	2	0.0032	Lyons	2	0.0064
Higgins	2	0.0032	Martin	2	0.0064
Howroyd	2	0.0032	Mckay	2	0.0064
Hudson	2	0.0032	Miller	2	0.0064
Hughes	2	0.0032	Newman	2	0.0064
Johnston	2	0.0032	Pearsall	2	0.0064
Total	98			87	

political families.

4.4 Implied Social Mobility Rates

Following Clark (2014), the method of estimating a social mobility rate in this paper is to assume that some group has some prevalence in generation t, x_t, and to see how that prevalence adjusts in generation $t+1$:

$$x_{t+1} = bx_t + e_t. \tag{4.1}$$

To construct the implied social mobility rates by this method, relative representations for a group of surnames need to be constructed. As specific genealogical data are not available, unusual surnames are used to establish the link between generations.

The collection of the unusual surnames drawn from the births dataset will be identified in the Hutchins School and parliament database and their relative representation determined when they exist. As the datasets for the late twentieth century are dated between 1980 and 2016, on average four 30-year periods, which will be used as the definition of a generation for this paper, will have elapsed before the present-day datasets. The relative representation of that collection can then be determined in the present-day datasets.

The first step to construct relative representations is to combine the counts of nineteenth century surnames for the various high-status groups. As the datasets are over time it is important to try to not double-count, for instance if a Hutchins School student later sat in parliament. To lessen concerns about this, the lists were compared and double-counting removed where it was identified.

The basis for inclusion in the rare-names, high-status, group is for there to be fewer than 100 counts of a particular surname in the births dataset, and three or more entries in the nineteenth century high-status group. On this basis there are 117 different surnames. Ten examples of these are: Giblin, Buscombe, Bisdee, Jeanneret, Bedford, Shoobridge, Finlayson, Crosby, Crowther, and Rockett. Of these 117 surnames, 22, or around 19 per cent, are also found in late twentieth century Tasmanian high-status surnames.

The relative representation of that group of 117 different surnames in the nineteenth century dataset is 11.89. As the relative representation is close to 12 this implies that group of surnames is about 12 times more represented in the high-status surnames than in the births dataset.

Keeping that same group of 117 different surnames, there are 26 of them in the present-day high-status group, which itself is of size 1,340. The broader population estimate is less well-established than it would be in comparable countries. The White Pages lists 571 households that had one of these 117 surnames, and there were 196,100 households in Tasmania in 2010. On the basis of this, the relative representation of this group in present-day Tasmania is 6.66.

If there are four generations between the present-day and nineteenth century datasets, then the implied estimate of b is 0.82. This estimate measures how similar the relative representation of names is in one generation compared with another, or the persistence of prevalence. Our estimate corresponds to both Clark (2014), who typically finds an estimate 'in the region 0.7–0.8' (p. 212) in various countries, and Clark and Cummins (2013) who find an estimate in the range of 0.73–0.9 in the U.K. over a longer time period. It is also similar to that of Clark et al. (2017) who applied this method in Australia with different data. More specifically, for instance, Clark et al. (2017) find a relative representation of rare surnames at universities of around 12 in 1900–1929, but a lower relative representation for doctors, where they find a value of around 2–3 in a similar time period. However, in both the cases of universities and doctors, the persistence that they find over time by comparing how these relative representations change is similar to our estimates. The advantage of our dataset compared with Clark et al. (2017) is that it covers a longer time period, however the disadvantage is that we are unable to use the electoral roll. Olivetti and Paserman (2015) finds a slightly lower income elasticity of around 0.3–0.5 in the U.S. between 1850 and 1930, using a slightly different method that is based on given names, rather than surnames.

4.4.1 Interpretation

Understanding what the 0.82 estimate (or Clark's usual 0.7–0.8 estimate) means can be difficult. The estimate is constrained to be positive, and if it were almost 0 then that would mean that few of the surnames that were prevalent in the previous generation were prevalent in the current one. If it were 1, then there would be essentially the same relative prevalence, and if it were, say, 2 then that would mean the relative prevalence of those surnames had increased. In that context, it is difficult to know whether a result of around 0.7–0.8 over the course of four generations should be concerning.

The main issue is the lack of a counterfactual. Although not a perfect approach, one way of going some way to address this is to randomly generate samples and compare the results in that context.[10] There are a variety of ways to use sampling to construct a counterfactual, but one way is to assume: there is some group of rare surnames of interest, S, (for instance, Giblin, Buscombe, ...); at the first generation there are N_1 people in the total population; n_1 people with a surname in S; H_1 people in the high-status group; and h_1 people in the high-status group with a surname in S.

In this set-up the observed proportion of surnames of interest in the total population is: $\frac{n_1}{N_1} = p_{1,N}$ and the observed proportion of surnames of interest in the high-status group is: $\frac{h_1}{H_1} = p_{1,H}$. So the relative prevalence is:

$$\frac{p_{1,H}}{p_{1,N}}.$$

Using the same notation for the fourth generation means that the estimate of b solves:

$$\frac{p_{1,H}}{p_{1,N}} \times b^3 = \frac{p_{4,H}}{p_{4,N}}.$$

Assume sampling with replacement for simplicity, then the number of people in the

[10] Thanks to Monica Alexander, Bruce Chapman, and participants at the UC Berkeley History Lunch, especially Martha Olney, who all independently suggested this approach, which is also similar to that of Olivetti and Paserman (2015).

high-status group with a surname in S is a draw from a binomial distribution:

$$h_1 \sim \text{Binomial}(H_1, p^*).$$

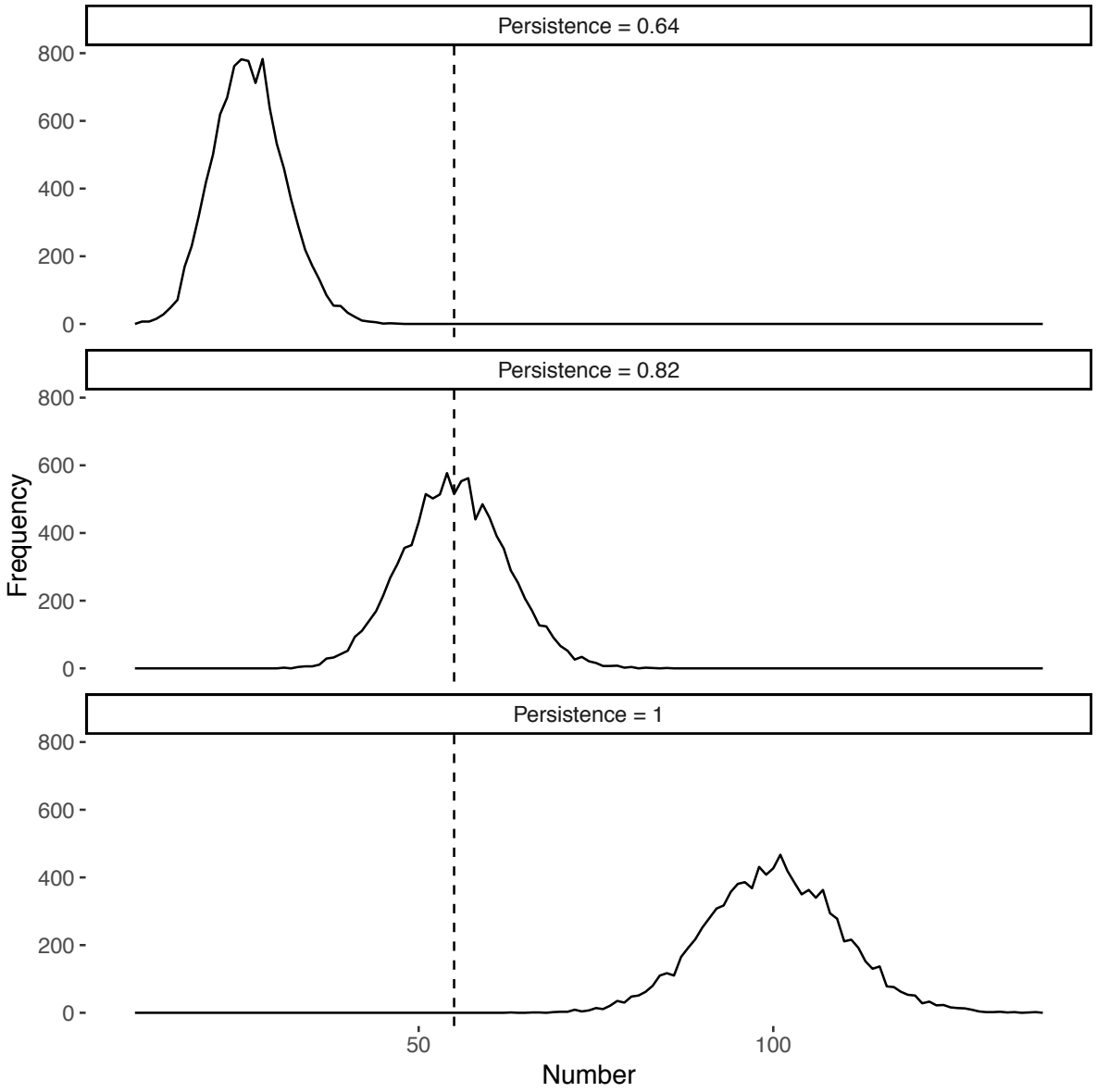
The true value of the p^* parameter is unknown, but the best estimate is $p_{1,H}$. Of interest for the purpose of constructing a counterfactual is the question of what if p^* was something else? In that situation, and with appropriate manipulation of p^* to account for generations, how likely is the number of observed surnames of interest in the high-status group after four generations, h_4?

Figure 4.8 shows the results of 10,000 draws from binomial distributions with different values of persistence. This is the implied number of those in the high-status group with a surname of interest after four generations if there were 1,000 people in the high-status group and 10,000 people in the population.

In this example, as can be seen in Figure 4.8, the true number of people in the high-status group with a surname of interest should be about 55 when persistence is 0.82, and the probability of obtaining a value between 50 and 60, inclusive, is a little over 0.55. When persistence decreases to 0.64, we expect a far fewer number of those in the high-status group with a surname of interest. And when persistence increases to 1, we expect far more of those in the high-status group with a surname of interest. However in both cases the probability of obtaining a value between 50 and 60 is essentially zero. The effect on the shape of the distribution is notable – as the number of people in the high-status group with a surname of interest must not be negative, the shape of the distribution becomes more bunched as persistence reduces.

In order for it to be likely to observe low values of persistence in this set-up, say less than 0.5, the number of individuals in the high-status group with a surname of interest would need to be less than around 25, as can be seen from Figure 4.8. At such low values, the distribution would be quite bunched, and a small change in the number of individuals would have an outsized effect on the persistence estimate. Additionally, when conducting these studies using real data, for privacy reasons or out of concern for drawing inference based on data errors, researchers may be hesitant to use an especially small number of

Figure 4.8: Impact of different persistence levels



individuals. But unless the number of individuals in the high-status group is substantially larger than 1,000, then for any reasonable number of individuals with a surname of interest the estimated persistence is likely to be in 0.7–0.8 range.

4.4.2 Robustness

The main driver of the estimated social mobility is the relative representation of the nineteenth century high-status group. One concern may be that this is an artefact of an aspect of the births dataset instead of underlying social mobility. To help lessen the

concern that the estimate of social mobility is due to a feature of the births dataset, the preceding analysis can be redone using the nineteenth century deaths dataset instead of the nineteenth century births dataset.

There are 94,603 deaths in the dataset. The same hurdle for inclusion can be used for the deaths dataset as was used for the births dataset, specifically, being so unusual as for there to be fewer than 100 counts of that surname in the entire deaths dataset, and that there is three or more entries of that surname in the nineteenth century high-status group. There are 169 different surnames that satisfy these criteria. All of the ten example surnames from before are also found here, apart from Jeanneret.

The relative representation of that group of 169 surnames in the nineteenth century dataset is 7.28. That is to say, those surnames in that group of high-status surnames are seven times as likely to be found in the high-status group than in the deaths dataset.

Of those 169 surnames, there are 95 instances of one of those surnames being in the late twentieth century Tasmanian high-status surname group. There are 1,340 surnames in the present-day Tasmanian high-status surname group. As such the relative representation of those surnames is 4.49. The implied social mobility rate is 0.91, which is slightly higher than that implied by the births dataset but not dissimilar.

Our main results were also robust to a sensitivity test in which we randomly removed 50 per cent of the surnames at a time and then re-ran the analysis. Although the specific relative representations did change, they changed in a way that was consistent over time and resulted in a similar estimate of persistence.

4.5 Conclusion

This paper is about social mobility, that is, how does the social status of one's ancestors affect one's own social status. The focus has been on Tasmania, for which relevant datasets are available for a little over 150 years. Although the datasets are not perfect, the estimated intergenerational persistence of social status is around 0.8. This is similar to other countries where the surname method has been used, and similar to other results in Australia using different datasets.

One of the weaknesses of this paper is the type of data used. For instance, a person is typically only awarded an Order of Australia when they are older, after they have distinguished themselves. As such, that dataset does not necessarily provide a contemporary measure of mobility in society. Another weakness in this paper is the patrilineal focus necessitated by using surnames. Some strategies for mitigating this are being developed, for instance Olivetti and Paserman (2015) are able to use the distribution of first names which they find convey information about socioeconomic status. Finally, migration presents a danger to the appropriateness of this approach. For instance, if the high-status group are less likely to emigrate from Tasmania, then the findings in this paper may reflect these different rates to a certain extent.

The strength of the approach in this paper is that by using the longer time frame allowed by surname analysis it should be less impacted by randomness. However, this means the research does not provide many strategies that policymakers could implement if they want to change the level of social mobility – few of us are in a position to pick our great-grandparents.

Possible extensions to this paper include improving the informational content of surnames, following Güell et al. (2014); expanding the geographic area of consideration to the whole of Australia; and improving the late twentieth century data. Although more work is needed in order to have more confidence in this estimate, the result suggests concern around economic inequality should remain at the centre of policy for some time to come.

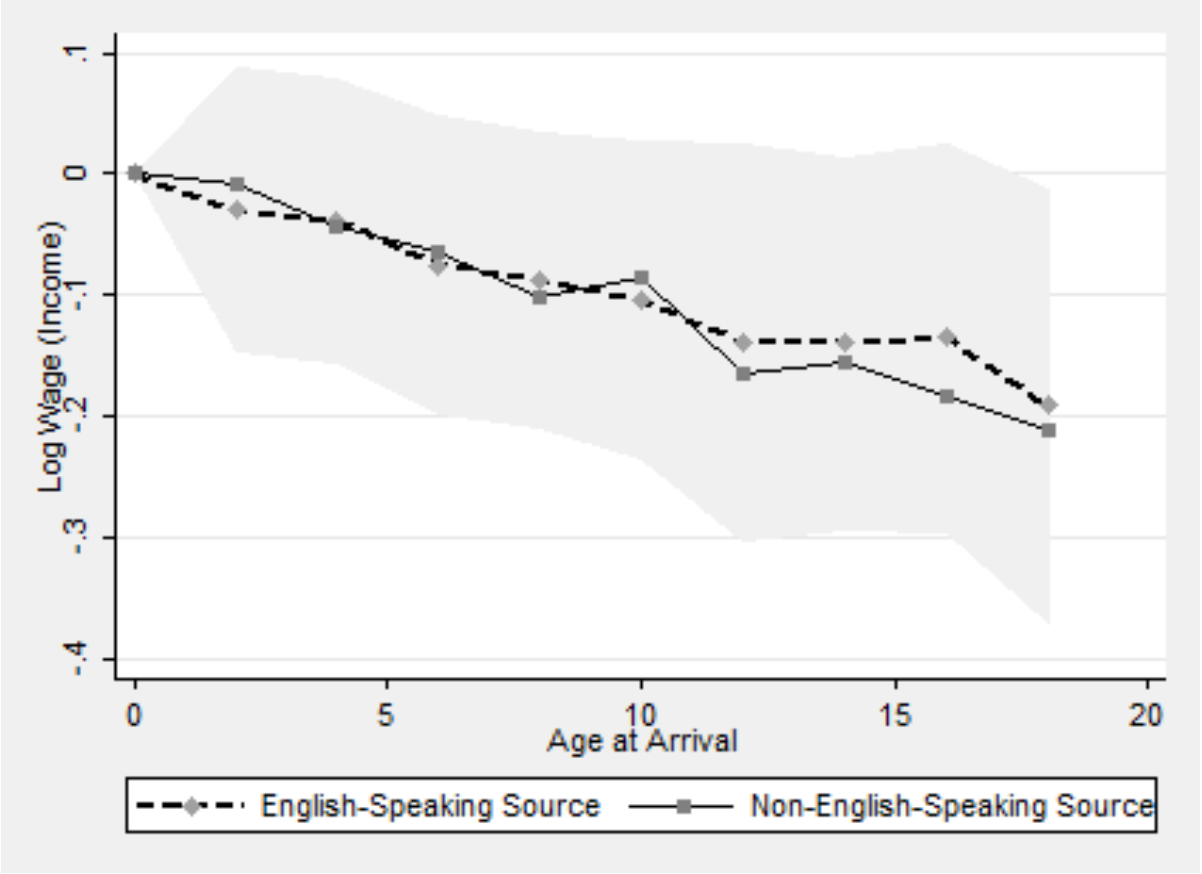
Appendices

Appendix A

Age-at-arrival appendices

A.1 Additional tables and graphs

Figure A.1: Age-at-arrival profile for English and non-English sources



Notes: Data is split by England, Scotland, Ireland and Wales versus all other sources. Sources: Linked sample of brothers from Ellis Island to the 1940 Census. Also see the text of the main document.

Table A.1: Age at arrival in census, 1899 to 1930 arrivals

Country of birth	Age at arrival	Percent 0–15 arrivals	Percent 16–45 arrivals	Percent 45+ arrivals
North and West Europe (old source)	22.7	24.6	71.2	4.1
South and East Europe (new source)	21.1	31.4	65.6	3.0
Russia	20.7	33.0	64.0	3.0
Romania	20.8	36.4	59.9	3.7
Portugal	20.8	33.5	63.4	3.1
Italy	21.4	31.9	64.7	3.4
Finland	21.5	19.6	78.8	1.6
Greece	21.6	27.7	69.9	2.4
Hungary	21.7	28.4	69.2	2.4
Netherlands	21.8	33.2	61.9	4.9
Austria	21.9	27.3	70.1	2.5
Norway	21.9	22.7	73.7	3.5
Sweden	22.0	21.0	75.8	3.2
Denmark	22.2	20.8	76.1	3.1
Spain	22.4	23.7	73.7	2.6
France	22.4	28.2	67.2	4.5
Ireland	22.4	19.8	76.5	3.7
Belgium	22.6	25.9	71.8	2.4
Scotland	23.1	27.2	67.3	5.5
Germany	23.5	24.1	70.9	5.0
England	23.6	27.6	66.4	6.0
Switzerland	23.7	19.7	76.3	4.0

Notes: Data is from the 1900 to 1930 United States Censuses, keeping only 1899 to 1930 arrivals. We keep these years to match with the years of arrival in Figure 2.1.
Source: See the text of the main document.

A.2 Further details on data creation

Information about Ellis Island arrivals was downloaded from `http://www.jewishgen.org/databases/EIDB/ellisgold.html`. The data collection focused on single males, aged 0–20, who arrived at Ellis Island between 1892 and 1924. The data fields that were collected were: first and last name; city and country of last residence; arrival day, month, and year; age at arrival; departure port; ship name; passenger id; and ethnicity. Sex and marital status were also collected, but just to restrict the sample to male and single. Passenger id is a unique identifier for each entry into Ellis Island and is numbered such that those next to each other on the ship manifest are next to each other for passenger

Table A.2: Robustness to higher quality links

	Income			Education		
Sample:	Main	High quality	Alternative match scores	Main	High quality	Alternative match scores
Age at arrival						
2 to 3	–0.0157	–0.0110	–0.0126	0.0229	0.0667	0.0325
	(0.0385)	(0.0523)	(0.0544)	(0.109)	(0.141)	(0.149)
4 to 5	–0.0421	–0.0234	–0.0392	–0.158	–0.165	–0.0794
	(0.0391)	(0.0520)	(0.0547)	(0.110)	(0.145)	(0.152)
6 to 7	–0.0680	–0.0889	–0.0659	–0.285	–0.325	–0.222
	(0.0407)	(0.0548)	(0.0566)	(0.115)	(0.151)	(0.160)
8 to 9	–0.0976	–0.114	–0.0907	–0.413	–0.407	–0.358
	(0.0406)	(0.0539)	(0.0578)	(0.116)	(0.151)	(0.162)
10 to 11	–0.0924	–0.102	–0.0874	–0.663	–0.602	–0.743
	(0.0429)	(0.0561)	(0.0621)	(0.123)	(0.159)	(0.173)
12 to 13	–0.159	–0.199	–0.121	–0.888	–0.802	–0.940
	(0.0487)	(0.0642)	(0.0692)	(0.143)	(0.185)	(0.209)
14 to 15	–0.150	–0.161	–0.158	–1.003	–0.909	–1.097
	(0.0502)	(0.0662)	(0.0728)	(0.145)	(0.187)	(0.210)
16 to 17	–0.168	–0.190	–0.177	–0.843	–0.744	–0.944
	(0.0536)	(0.0689)	(0.0787)	(0.153)	(0.198)	(0.227)
18 to 20	–0.204	–0.219	–0.117	–0.795	–0.675	–0.854
	(0.0587)	(0.0763)	(0.0946)	(0.167)	(0.220)	(0.263)
Observations	35,978	16,955	14,968	51,591	24,057	21,443
R-squared	0.659	0.634	0.694	0.632	0.628	0.668

Notes: Data is a sample of brothers linked from Ellis Island records to the 1940 Census. High-quality links are determined to be in the better 50 percent of scores for our linked dataset, as determined by the sum of Jaro-Winkler distance in first name, Jaro-Winkler distance in last name, and absolute difference in year of birth. The excluded group is arrivals at age zero and one. Brothers fixed effects are included in each column. Standard errors are clustered by household.
Source: See the text of the main document.

id. Since families are listed together in ship manifests, we can identify brothers as those listed next to each other who have the same surname, after sorting by ship name and passenger id. Since we do not collect females, we still capture brothers even if brothers were not immediately next to each other on the original manifests because the brothers appear next to each other in our data of only males. This leaves us with 447,540 potential brothers.

Next, we clean the residence field. From that field we needed a city and a country of origin; however, the initial origin field needs a significant amount of cleaning. For instance, the field contains abbreviations, inconsistent spelling, and differing amounts of information (for instance, just the city name, or the city name and the state, or the city

Table A.3: Effect of age at arrival on labour supply, weekly wages, and self-employment

Age at arrival:	LFP	Weeks of work	Log (Weekly Wage)	Self employed	Self em. and not farmer
2 to 3	–0.00181	0.188	–0.0130	–0.0312**	–0.0170
	(0.00918)	(0.641)	(0.0291)	(0.0151)	(0.0140)
4 to 5	–0.00434	–0.552	–0.0201	–0.0118	0.00648
	(0.00910)	(0.647)	(0.0291)	(0.0152)	(0.0141)
6 to 7	0.000293	–0.388	–0.0385	–0.0304*	0.00122
	(0.00946)	(0.670)	(0.0304)	(0.0159)	(0.0148)
8 to 9	–0.00386	–0.668	–0.0614**	–0.0302*	0.00279
	(0.00949)	(0.684)	(0.0307)	(0.0159)	(0.0149)
10 to 11	–0.00164	–0.481	–0.0739**	–0.0392**	–0.00152
	(0.00998)	(0.714)	(0.0328)	(0.0169)	(0.0156)
12 to 13	0.0132	0.00527	–0.107***	–0.0367*	0.00714
	(0.0110)	(0.805)	(0.0375)	(0.0194)	(0.0181)
14 to 15	0.00231	–0.891	–0.0965**	–0.0372*	0.00590
	(0.0117)	(0.836)	(0.0382)	(0.0199)	(0.0185)
16 to 17	–0.0106	–1.178	–0.138***	–0.0538**	0.00782
	(0.0125)	(0.885)	(0.0405)	(0.0211)	(0.0198)
18 to 20	–0.00472	–1.297	–0.158***	–0.0481**	0.0131
	(0.0140)	(0.959)	(0.0454)	(0.0235)	(0.0219)
Observations	53,129	53,129	35,663	47,901	47,901
R-squared	0.480	0.489	0.655	0.559	0.546

Notes: *** p<0.01, ** p<0.05, * p<0.10. Data is a sample of brothers linked from Ellis Island records to the 1940 Census. The number of observations changes across columns because only wage workers are included in the third column, and those who have missing information from the self-employed category are dropped in the fourth column. The excluded group is arrivals at age zero and one. Brothers fixed effects are included in each column. Standard errors are clustered by household.
Source: See the text of the main document.

name, state, and country). We clean the country of birth variable for origins that have ten or more observations, or 319,510 of our initial sample of 447,540 brothers.[1] For records that do not identify the country of birth, we assume the country of birth based on the reported ethnicity. This is mostly straightforward, but we cannot match for ethnicities such as Jewish, Arabian, or Black. Most of the time there is a second ethnicity listed for these sources, but if not, then we dropped those (12,620 observations) from our dataset.

Next, we also wish to link on year of birth, but the Ellis Island records only have age and date of arrival rather than year of birth. Therefore, we need to back out year of birth,

[1]Of the approximately 320,000 observations that have more than 10 entries, about 0.5 percent of countries of origin could not be identified. We assume that the country of origin matches one's ethnicity.

Table A.4: Effect of age at arrival on home ownership and location

Age at arrival:	Own house	Log(value of house)	Urban	Urban population
2 to 3	0.00106	–0.00147	–0.00870	281.8
	(0.0179)	(0.117)	(0.0130)	(318.3)
4 to 5	0.0102	0.00106	–0.0102	568.7*
	(0.0182)	(0.120)	(0.0131)	(322.0)
6 to 7	0.0116	–0.0261	0.00170	459.4
	(0.0188)	(0.126)	(0.0137)	(334.9)
8 to 9	0.0190	–0.0525	–0.00727	427.7
	(0.0189)	(0.125)	(0.0138)	(337.0)
10 to 11	0.0236	–0.0796	–0.0114	622.2*
	(0.0199)	(0.128)	(0.0145)	(353.9)
12 to 13	–0.00468	0.0291	–0.00791	549.8
	(0.0227)	(0.141)	(0.0164)	(400.9)
14 to 15	0.0217	–0.0639	0.00208	478.7
	(0.0230)	(0.142)	(0.0167)	(411.4)
16 to 17	0.0207	–0.0379	–0.00851	937.5**
	(0.0242)	(0.153)	(0.0176)	(430.9)
18 to 20	0.0208	–0.0772	–0.00479	769.8
	(0.0265)	(0.168)	(0.0193)	(469.2)
Observations	51,616	20,746	53,129	53,129
R-squared	0.521	0.788	0.572	0.581

Notes: *** p<0.01, ** p<0.05, * p<0.10. Data is a sample of brothers linked from Ellis Island records to the 1940 Census. The number of observations changes across columns because missing information is dropped, and only those who own a house are in the second column. The excluded group is arrivals aged zero and one. Brothers fixed effects are included in each column. Standard errors are clustered by household.
Source: See the text of the main document.

which is typically done with the formula: Year of Observation-Age. However, this implies that an arrival who listed their age as 10 and arrived on 1 January 1910 would be born in 1900, but this arrival instead was likely born in 1899. Therefore, we back out the year of birth as Year of Arrival - Age for those who arrived in the second half of the year, and Year of Arrival-Age-1 for those who arrived in the first half of the year.

Third, we drop those who have missing letters in their first or last names, which is identified by strings such as "..." or "?," which drops 4,653 individuals.

Fourth, if an individual lists an initial as the first name, but then a longer second name, then we keep the second name as the main name; however, we drop those who only report an initial for the first name and give no second name.

Table A.5: Robustness when controlling for birth order

Age at arrival:	Income	Income	Income	Education	Education	Education
2 to 3	–0.0157	–0.0215	–0.0202	0.0229	0.0488	0.0511
	(0.0385)	(0.0398)	(0.0398)	(0.109)	(0.113)	(0.113)
4 to 5	–0.0421	–0.0539	–0.0523	–0.158	–0.107	–0.105
	(0.0391)	(0.0437)	(0.0437)	(0.110)	(0.124)	(0.124)
6 to 7	–0.0680*	–0.0845*	–0.0831*	–0.285**	–0.214	–0.213
	(0.0407)	(0.0491)	(0.0491)	(0.115)	(0.139)	(0.139)
8 to 9	–0.0976**	–0.119**	–0.118**	–0.413***	–0.322**	–0.320**
	(0.0406)	(0.0536)	(0.0536)	(0.116)	(0.154)	(0.154)
10 to 11	–0.0924**	–0.119*	–0.118*	–0.663***	–0.550***	–0.547***
	(0.0429)	(0.0613)	(0.0613)	(0.123)	(0.175)	(0.175)
12 to 13	–0.159***	–0.189***	–0.188***	–0.888***	–0.758***	–0.756***
	(0.0487)	(0.0695)	(0.0694)	(0.143)	(0.202)	(0.203)
14 to 15	–0.150***	–0.185**	–0.184**	–1.003***	–0.854***	–0.849***
	(0.0502)	(0.0764)	(0.0764)	(0.145)	(0.217)	(0.217)
16 to 17	–0.168***	–0.207**	–0.205**	–0.843***	–0.675***	–0.665***
	(0.0536)	(0.0838)	(0.0837)	(0.153)	(0.238)	(0.238)
18 to 20	–0.204***	–0.252***	–0.248**	–0.795***	–0.589**	–0.559**
	(0.0587)	(0.0967)	(0.0967)	(0.167)	(0.279)	(0.280)
Birth order linear:						
Birth order		–0.0121			0.0522	
		(0.0195)			(0.0558)	
Birth order dummies:						
Second born			–0.00879			0.0902
			(0.0213)			(0.0613)
Third born			–0.0429			0.0898
			(0.0453)			(0.130)
Fourth born			–0.00381			–0.138
			(0.0786)			(0.221)
Observations	35,976	35,976	35,976	51,591	51,591	51,591
R-squared	0.659	0.659	0.659	0.632	0.632	0.632

Notes: *** p<0.01, ** p<0.05, * p<0.10. Data is from a sample of brothers linked from Ellis Island records to the 1940 Census. All regressions control for sibling fixed effects. The excluded group is arrivals at age zero and one. Brothers fixed effects are included in each column. Standard errors are clustered by household.
Source: See the text of the main document.

Fifth, we Americanise the names. The first names found in the data were anglicized to increase the likelihood of matching. For instance "Giuseppe" was changed to "Joseph." Each name was run through `behindthename.com` to provide a list of related names. To anglicize a name required at most a many-to-one relationship between the original name and the anglicized one. The issue was that this was found to be a many-to-many mapping; for instance, Joseph maps to both Joe and Guiseppe, but both of those also map back to

Table A.6: Robustness to dropping arrivals 16 years and older

Age at arrival	Income	Income	Education	Education
2 to 3	–0.0157	–0.0155	0.0229	0.0250
	(0.0385)	(0.0396)	(0.109)	(0.113)
4 to 5	–0.0421	–0.0419	–0.158	–0.159
	(0.0391)	(0.0401)	(0.110)	(0.115)
6 to 7	–0.0680*	–0.0667	–0.285**	–0.297**
	(0.0407)	(0.0419)	(0.115)	(0.120)
8 to 9	–0.0976**	–0.0992**	–0.413***	–0.415***
	(0.0406)	(0.0419)	(0.116)	(0.122)
10 to 11	–0.0924**	–0.0968**	–0.663***	–0.667***
	(0.0429)	(0.0446)	(0.123)	(0.129)
12 to 13	–0.159***	–0.164***	–0.888***	–0.933***
	(0.0487)	(0.0516)	(0.143)	(0.154)
14 to 15	–0.150***	–0.157***	–1.003***	–1.066***
	(0.0502)	(0.0547)	(0.145)	(0.161)
16 to 17	–0.168***		–0.843***	
	(0.0536)		(0.153)	
18 to 20	–0.204***		–0.795***	
	(0.0587)		(0.167)	
Observations	35,976	28,982	51,591	40,837
R-squared	0.659	0.675	0.632	0.657

Notes: *** p<0.01, ** p<0.05, * p<0.10. Data is from a sample of brothers linked from Ellis Island records to the 1940 Census. All regressions control for sibling fixed effects. The excluded group is arrivals at age zero and one. Brothers fixed effects are included in each column. Standard errors are clustered by household. Columns 2 and 4 drop those who arrived older than age 16.
Source: See the text of the main document.

Joseph. This meant that the mapping was circular and would depend on the order that the names were processed. Additionally, it was not clear from `behindthename.com` which name was best considered the anglicized version—should Joseph be changed to Guiseppe or vice versa?

To address these issues the database of first names at birth from U.S. censuses that occurred before 1930 were obtained and combined to give a ranking of the popularity of each first name, as defined by the number of U.S. born children with that name. For each mapping, grouped by the initial name, say Guiseppe to Guisep and Guiseppe to Joseph, the script provided a preferred choice, based on which of the possible names is the most popular in the U.S. census dataset. With this, we created a data file that included two

Table A.7: Age-at-arrival profiles are robust to Americanisation process

	Income		Education	
Linking:	Main	Non-Americanised	Main	Non-Americanised
Age at arrival:				
2 to 3	–0.0157	–0.0419	0.0229	–0.131
	(0.0385)	(0.0536)	(0.109)	(0.147)
4 to 5	–0.0421	–0.0254	–0.158	–0.280*
	(0.0391)	(0.0529)	(0.110)	(0.152)
6 to 7	–0.0680*	–0.0511	–0.285**	–0.360**
	(0.0407)	(0.0566)	(0.115)	(0.158)
8 to 9	–0.0976**	–0.102*	–0.413***	–0.670***
	(0.0406)	(0.0552)	(0.116)	(0.157)
10 to 11	–0.0924**	–0.0364	–0.663***	–0.774***
	(0.0429)	(0.0600)	(0.123)	(0.169)
12 to 13	–0.159***	–0.131*	–0.888***	–1.178***
	(0.0487)	(0.0693)	(0.143)	(0.197)
14 to 15	–0.150***	–0.169**	–1.003***	–1.326***
	(0.0502)	(0.0746)	(0.145)	(0.199)
16 to 17	–0.168***	–0.168**	–0.843***	–1.012***
	(0.0536)	(0.0769)	(0.153)	(0.215)
18 to 20	–0.204***	–0.201**	–0.795***	–1.062***
	(0.0587)	(0.0836)	(0.167)	(0.235)
Observations	35,977	18,052	51,591	25,712
R-squared	0.659	0.678	0.632	0.662

Notes: *** p<0.01, ** p<0.05, * p<0.10. Data is from a sample of brothers linked from Ellis Island Records to the 1940 Census. This table tests the robustness of results when not Americanising names in our dataset. See Appendix C for more detail. All regressions control for sibling fixed effects. The excluded group is arrivals at age zero and one. Brothers fixed effects are included in each column. Standard errors are clustered by household. Source: See the text of the main document.

primary variables: the first name string as observed in the Ellis Island name, and the Americanised name. We then merged our Ellis Island dataset with this file to attach the Americanised name to our dataset.

Finally, we drop potential brothers who are next to each other and are more than ten years apart. We do this in case those with the same surname that are more than ten years apart are not truly brothers, but represent a father-son relationship or uncle-nephew relationship. Note that this does not drop sets of brothers where the oldest and youngest are more than ten years apart. For example, if there are a 14, five, and one year-old who are identified as potential brothers, we keep them since none are more than ten years

apart; but if there is a 20 year old, five, and one year-old, we drop the 20 year old from the dataset. Keeping those more than ten years apart does not lead to a qualitative change in results. Ultimately, we are left with 397,137 potential brothers to link.

A.3 Linking methodology

We link our cleaned dataset of 397,137 brothers to white males in the 1940 U.S. Census by searching for the best match among the potential set of matches on Americanised first name, last name, year of birth (within a range of three years), and country of birth. Our process follows the same idea as others in the literature (Abramitzky et al., 2014) with a few modifications. The main difference in our methodology is that we first Americanise all foreign-born names in the 1940 Census in case an immigrant changed his name from, for example, Jörg to George. Another difference is that we rate the quality of potential matches by determining the differences in string similarity via the Jaro-Winkler algorithm. The steps to our linking process are as follows:

- "Americanise" the first names of Ellis Island and census records with a list of 28,000 name variants from `behindthename.com`. Names that do not have an American equivalent are unchanged.
- Standardise the first name resulting from step one and the last name with the NYSIIS algorithm. Drop observations that have the same Americanised first name string, last name string, year of birth, and country of birth in both the Ellis Island records and 1940 Census.
- Find all possible matches on NYSIIS Americanised first name, NYSIIS last name, country of birth, and exact year of birth. Repeat this step, but expand the window for difference in year of birth to allow up to a three-year difference.
- Calculate a match score for each potential match, which is the sum of the Jaro-Winkler distance in Americanised first name string, Jaro-Winkler distance in last name string, and difference in year of birth (zero for exact year of birth match).

Note that this method does not actually treat all NYSIIS names equally, but only uses the NYSIIS algorithm to find potential matches.

- Keep the minimum match score for each observation in the Ellis Island records, and then keep the minimum match score for each observation in the 1940 Census.

This process leads to linking 103,005 individuals from the set of 397,137 brothers in the Ellis Island records, a match rate of 25.9 percent, a reasonable rate for a single match. Since the empirical strategy requires the use of siblings, we drop individuals who do not have another matched sibling, which leads to our final sample of 53,129 brothers used in the main text.

In Table A.8, we show differences between our entire linked sample and the sample of brothers we use in the main text. The primary difference between samples is that people in the brothers' sample are 12 percentage points less likely to be from Southern and Eastern Europe, which is unsurprising since these sources had lower linking rates and thus there are fewer sets of two brothers linked than single individuals linked. This also leads our brothers sample to be slightly higher skilled than the overall linked sample by 0.2 years of education and by earning 2.9 percent more wage income.

While it is well known that linking may bias the representativeness of the sample, we cannot directly test the representativeness because the 1940 Census does not include year of arrival, and thus we cannot compare our sample to those from the same arrival cohort and those with the same arrival age. However, we can show how our linked dataset of brothers has different attributes than the European migrant stock with the same years of birth in the 1940 Census. The difference is shown in Columns III and IV in Table A.8; note that differences between our linked sample of brothers and the 1940 migrant stock may result from biases in the linking process, because we have a specific migrant cohort, or because we only keep those who arrived at young ages. As expected, our linked sample of brothers is higher skilled and earns more income than the 1940 Census as a whole, partially because we have younger arrivals and younger arrivals have higher earnings later in life.

One way in which our sample may be unrepresentative is because we Americanise

Table A.8: Characteristics of our linked sample of brothers in the 1940 census

Sample:	I Linked sample	II Non-brothers	III Brothers	IV 1940 Stock	(III–II) Difference	(III–VI) Difference
Age	40.88	40.84	40.91	48.04	0.0677	–7.123
	(9.040)	(8.913)	(9.152)	(11.56)	(0.0571)	(0.0654)
Education	7.171	7.051	7.277	6.860	0.227	0.417
	(3.705)	(3.762)	(3.650)	(3.951)	(0.0238)	(0.0241)
Log occ.	6.902	6.879	6.924	6.910	0.0451	0.0137
score	(0.335)	(0.327)	(0.341)	(0.330)	(0.00218)	(0.00223)
Self-employed	0.201	0.210	0.193	0.233	–0.0170	-0.0393
	(0.401)	(0.408)	(0.395)	(0.422)	(0.00267)	(0.00273)
Log income	6.951	6.936	6.965	6.887	0.0295	0.0778
wage	(0.709)	(0.711)	(0.707)	(0.777)	(0.00547)	(0.00581)
South and	0.642	0.712	0.580	0.617	–0.132	–0.0372
East Europe	(0.479)	(0.453)	(0.494)	(0.486)	(0.00300)	(0.00309)
Age arrival	2.332	0	4.412			
Diff. in family	(3.194)	(0)	(3.181)			
N	100,476	47,353	53,123	47,667		

Notes: This table shows descriptive statistics of: the linked sample; the linked sample split into brothers and non-brothers; and then the 1940 Census. Note that not all individuals have observed wage income, years of education, or self-employment status.
Source: Linked sample of brothers from Ellis Island records to 1940 Census, and also 1940 Census (Ruggles et al., 2017).

names and this introduces a bias in our linking process. In our dataset, 36 percent of matches are matched due to the Americanisation process. Given that about 30 percent of immigrants switched their first names at the naturalization stage according to data from New York, and that arrival records had more foreign-sounding names than census records, we believe that 36 percent is a reasonable number (Biavaschi et al. (2017); Carneiro et al. (2015)). In Table A.9, we list the top 25 names that were Americanised in our dataset of linked brothers. At the top of the list are primarily Italian names such as Giuseppe, the alternative (and misspelled) Guiseppe, Giovanni, and Antonio. There are also non-Italian names that are Americanised, such as Josef, Johann, and Wilhelm.

Americanising names is a not a standard process when linking individuals and therefore may somehow drive our results. We perform a robustness check in which we link the arrival records with the 1940 U.S. Census without Americanising any of the names in the Ellis Island records or the 1940 Census. Not Americanising names leads to a smaller set of

Table A.9: Top 25 Americanisations in linked dataset of brothers

Rank	First Name: Arrival	First Name: 1940	N
1	Giuseppe	Joseph	2,227
2	Giovanni	John	1,567
3	Antonio	Anthony	1,362
4	Luigi	Louis	860
5	Vincenzo	Vincent	858
6	Guiseppe	Joseph	835
7	Pietro	Peter	715
8	Michele	Michael	585
9	Josef	Joseph	545
10	Domenico	Dominick	453
11	Jan	John	439
12	Nicola	Nicholas	299
13	Paolo	Paul	257
14	Johann	John	240
15	Carlo	Charles	185
16	Johan	John	183
17	Wilhelm	William	178
18	Johannes	John	164
19	Jose	Joseph	162
20	Heinrich	Henry	159
21	Andrea	Andrew	155
22	Janos	John	151
23	Georg	George	142
24	Filippo	Philip	141
25	Tommaso	Thomas	128

Notes: This table lists the top 25 Americanisations in our linked dataset of brothers, where the arrival name is the one listed in the Ellis Island records, while the 1940 name is the one listed in the 1940 U.S. Census.
Source: Linked sample of brothers from Ellis Island records to 1940 Census.

linked individuals of 67,427, a drop of about 30 percent. This leads to an even smaller set of two successfully linked brothers of 26,412, which is unsurprising since we do not link those who changed their name. The smaller set of observations leads to noisier estimates, but our qualitative results hold when using the non-Americanised dataset, suggesting that the Americanisation process does not drive the results in the main text. Table A7 shows the results for log wage income and years of education when not Americanising our data compared with our main results.

A.4 Creation of immigrant-specific occupational score

In this section, we provide further details on the creation of the immigrant-specific occupational score used in text. We create this score to improve on the standard occupational scores used in the literature, such as the 1950 occscore from IPUMS and the 1901 Cost of Living Survey score. There are important limitations when using these commonly used scores; for example, the 1901 Cost of Living Score is only representative for married urban families and therefore does not provide an accurate estimate for rural or single workers. The 1950 occupational score reflects earnings after WWII, and therefore understates wage gaps for data prior to WWII (Goldin and Margo, 1992). Moreover, neither score reflects earnings that are specific to immigrants and thus they understate any difference between immigrants and natives, a key interest for this paper.

We create an alternative occupational score that is based on income reported in the full-count 1940 U.S. Census. Our approach follows Collins and Wanamaker (2014) and Collins and Wanamaker (2017) in that we impute income separately by group; but instead of groups separated by race and region as in Collins and Wanamaker (2017), we impute income separately by country of birth. Therefore, the occupation score is essentially the average earnings in each occupation/country of birth cell. We provide further details on how we create the score below, but we follow Appendix I.b of Collins and Wanamaker (2017) to fix for self-employed earnings and non-monetary compensation for farm laborers and farmers.

First, we take the full-count 1940 U.S. Census and top-code income to $5,000 for wage workers. For self-employed workers, we ignore their reported wage income since this is not consistently reported, but we instead impute their income. To do this, we follow the strategy laid out by Collins and Wanamaker (2017) where we take the ratio of self-employed earnings to wage-worker earnings by occupation in the 1960 census, assume this ratio from 1960 is a good proxy for the ratio in 1940, and multiply the ratio with the mean wage income by occupation and country of birth. This leads to an imputed income for each self-employed person that varies by occupation and country of birth. Then we collapse the 1940 data by detailed occupation code and country of birth to get an average

income for each occupation, which forms the occupational score for the large majority of our data.

We do not take the above approach for farm laborers and farmers because they may receive compensation in kind which is not recorded in the income data. We take a few extra steps to estimate their incomes. Starting with farm laborers and once again following Collins and Wanamaker (2017), we increase farm laborers' mean wage income in the 1940 Census by 26 percent to reflect in-kind compensation, which is based on the 1957 USDA report Major Statistical Series of the U.S. Department of Agriculture. The next step is to estimate income for farmers. First, we assume that the perquisite rate of farmers in the 1960 Census is 35 percent (also based on the USDA report), and we scale up their reported (wage and business) income by this factor. To create the final estimate for farmer income in 1940, we assume that the ratio between farm laborers and farmer income (inclusive of perquisites) in 1960 is the same as in 1940. Therefore, we need to estimate farm laborers' income in 1960, which we boost their income by 19 percent to reflect in-kind compensation.

A.5 Robustness of results to a linking approach related to Feigenbaum (2016)

A.5.1 An alternative approach to linking

In this section, we discuss an alternative method of linking immigrants from Ellis Island records to the 1940 Census that is related to Feigenbaum (2016). In the main text, our method of picking the best link is based on Massey (2017) where we rate matches by summing the difference in year of birth, Jaro-Winkler distance in first name, and Jaro-Winkler distance in last name. Rather than rating matches based on these values, we could instead use training data to estimate the penalty for having deviations in year of birth, first name and last name, as well as other variables. Feigenbaum (2016) uses this approach when linking children from the 1915 Iowa Census to the 1940 U.S. Census.

Related to our study on immigrants, Ward (2018) applies the Feigenbaum (2016) method to immigrants during the Age of Mass Migration in his study of English fluency in the 1910 to 1930 Censuses. Concerned that the penalty for deviations in name may vary by the source of immigrants, Ward (2018) draws random samples of 2,000 from 16 different ethnicities in 1920 (e.g., Polish, Italian, German, etc.), hand-links them to the 1930 U.S. Census, and estimates a model to predict a match score for each immigrant.[2] We use Ward (2018) hand-linked data on immigrants between the 1920 and 1930 Censuses as "training data" for our sample of Ellis Island arrivals linked to the 1940 Census. While linking Ellis Island records to the 1940 Census is different than linking the 1920 Census to 1930 Census, it will serve as a good quality check on our main results in sample. We do not use the data created from this linking process as our main sample because the "training data" is not specific to our linked data between Ellis Island arrival records and the 1940 Census; however, qualitative results from this dataset are consistent with results in the main paper.

We cannot directly use the estimated probit coefficients from Ward (2018) since he predicts scores based on year of arrival, a variable that is unavailable in the 1940 Census; therefore, we re-estimate a probit for each of the 16 ethnicities after removing the year of arrival variables from the model. The results for each probit model are shown in Tables A.10–A.13, and show generally that having smaller deviations in Jaro-Winkler distance and year of birth predicts a match. We can then use the coefficients from this model to predict the probability that each potential link would be a match. Potential links between the Ellis Island data and the 1940 Census are chosen such that they have a Jaro-Winkler distance in first name of less than 0.20, Jaro-Winkler distance in last name of less than 0.25, a year of birth distance of less than 3, an exact match on country of birth, the same first letter of the first name, and the same first letter of the last name.

At this point, we have predicted match probabilities for each potential link; now we

[2] Ward (2018) discusses linking 15 different ethnicities (these are described as ethnicities in the original Ellis Island datasource, however they are arguably nationalities): German, Jewish, Dutch, Swedish, Danish, Norwegian, Italian, French, Romanian, Greek, Russian, Czech/Slovak, Polish, Finnish, and Hungarian. Ward (2018) additionally links immigrants from English-speaking sources (that is, England, Ireland, and Scotland), but does not report this since his study is on the acquisition of English skills for immigrants from non-English-speaking sources.

have to determine parameters for who is included in the dataset. We choose the meta-parameters shown in Table A.14 following the conservative strategy of Ward (2018) such that the PPV (predictive positive value, or estimated share of true positives to overall positives) is 0.90.[3] This method of being conservative to increase the number of true positives (or reduce the number of false positives) leads to a significantly lower linking rate than the training sample and compared to our main method of linking immigrants.

This linking process leads to a smaller sample of brothers of 21,994 compared with our main sample of 53,129. This is partially because it is difficult to predict the best link from observable variables in hand-linked data; it also may be because the data we use to predict links is not specific to the Ellis Island records matched to the 1940 Census. Being more restrictive about who is kept in the sample may also lead to biases in representativeness, but once again this is difficult to determine given the lack of census or representative sample that observes immigrant outcomes in 1940 in addition to year/age of arrival. We follow the same weighting process in the main section and weight to ensure that our sample is representative on country of birth.

A.5.2 All results qualitatively hold with alternative linked sample

We recreate all tables and figures from the main text with this linked sample (see Tables A.15–A.17; Figures A.2–A.4). We show that all results are qualitatively the same as in the main text: the age-at-arrival and income profile are similarly sloped with or without brothers fixed effects, older arrivals experienced a larger native-immigrant wage gap than younger arrivals, older arrivals acquired fewer years of education than younger arrivals, and older arrivals were less likely to marry a native-born spouse.

[3] The meta-parameters are the cut-off of predicted probability for keeping an immigrant in the sample, and the minimum ratio between highest match score and second-highest match score (to drop close second matches).

Table A.10: Probit coefficients, Part 1

	English	German	Yiddish, Jewish	Dutch
Year of birth difference = 1	–0.728***	–0.454***	–0.738***	–0.782***
	(0.0825)	(0.0865)	(0.0743)	(0.103)
Year of birth difference = 2	–1.104***	–0.705***	–0.974***	–1.201***
	(0.0962)	(0.0986)	(0.0842)	(0.135)
Year of birth difference = 3	–1.163***	–1.025***	–1.180***	–1.395***
	(0.104)	(0.116)	(0.0986)	(0.146)
Jaro-Winkler distance in first name string	–6.630**	–2.495	–3.574	1.330
	(2.892)	(2.366)	(2.611)	(2.029)
Jaro-Winkler distance in last name string	–9.190***	–9.705***	–8.584***	–12.13***
	(0.848)	(0.723)	(0.756)	(0.925)
Exact first name match (NYSIIS)	-0.204	0.102	0.0417	0.601*
	(0.404)	(0.313)	(0.347)	(0.314)
Exact first and last name match (NYSIIS)	–0.342***	–0.401***	–0.280***	–0.533***
	(0.105)	(0.120)	(0.0926)	(0.173)
Total number of hits	–0.171***	–0.202***	–0.165***	–0.266***
	(0.0190)	(0.0189)	(0.0220)	(0.0246)
Total number of hits squared	0.00413***	0.00521***	0.00347***	0.00653***
	(0.000655)	(0.000673)	(0.000717)	(0.000916)
First letter of last name match	0.230	-0.116	0.121	–0.487***
	(0.173)	(0.119)	(0.153)	(0.167)
First letter of first name match	0.171	0.557***	1.596***	0.385
	(0.291)	(0.181)	(0.528)	(0.262)
More than two hits have NYSIIS last name match	0.533***	0.535***	0.636***	–1.041***
	(0.129)	(0.165)	(0.116)	(0.266)
One hit has NYSIIS last name match	1.223***	0.848***	1.383***	1.958***
	(0.126)	(0.163)	(0.119)	(0.253)
Jaro-Winkler distance in NYSIIS first name	–1.548**	–3.213***	–2.540***	–5.875***
	(0.758)	(0.637)	(0.719)	(0.949)
Jaro-Winkler distance in NYSIIS last name	–1.255**	–0.308	–0.0312	–0.534*
	(0.584)	(0.257)	(0.122)	(0.319)
Middle initial match, if have one	1.116***	1.624***	0.414	1.011***
	(0.126)	(0.318)	(0.727)	(0.315)
Constant	0.837	1.133***	–0.625	2.500***
	(0.530)	(0.408)	(0.672)	(0.474)
Observations	12,975	11,227	25,691	6,651

Source: Ward (2018).

Table A.11: Probit coefficients, Part 2

	Swedish	Danish	Norwegian	Italian
Year of birth difference = 1	–0.824***	–0.722***	–0.685***	–0.445***
	(0.0693)	(0.0733)	(0.0822)	(0.0652)
Year of birth difference = 2	–1.060***	–1.140***	–1.161***	–0.822***
	(0.0803)	(0.0916)	(0.105)	(0.0808)
Year of birth difference = 3	–1.342***	–1.446***	–1.459***	–1.212***
	(0.0973)	(0.114)	(0.123)	(0.107)
Jaro-Winkler distance in first name string	–4.822***	–4.209***	–2.601*	0.906
	(1.466)	(1.552)	(1.512)	(1.243)
Jaro-Winkler distance in last name string	–6.467***	–5.573***	–7.379***	–10.49***
	(0.828)	(0.889)	(0.793)	(0.592)
Exact first name match (NYSIIS)	0.228	0.268	0.403*	0.462**
	(0.226)	(0.217)	(0.222)	(0.188)
Exact first and last name match (NYSIIS)	–0.0560	–0.0728	–0.361***	0.00105
	(0.0954)	(0.0947)	(0.107)	(0.0932)
Total number of hits	–0.175***	–0.218***	–0.215***	–0.0654***
	(0.0182)	(0.0200)	(0.0194)	(0.0246)
Total number of hits squared	0.00366***	0.00485***	0.00472***	0.000384
	(0.000608)	(0.000670)	(0.000684)	(0.000787)
First letter of last name match	0.217	0.464**	0.354**	–0.00628
	(0.133)	(0.182)	(0.150)	(0.143)
First letter of first name match	0.446***	0.956***	0.740***	0.151
	(0.158)	(0.198)	(0.187)	(0.107)
More than two hits have NYSIIS last name match	0.690***	0.0203	–0.0275	0.686***
	(0.121)	(0.153)	(0.156)	(0.118)
One hit has NYSIIS last name match	1.229***	1.389***	1.547***	0.713***
	(0.128)	(0.157)	(0.153)	(0.123)
Jaro-Winkler distance in NYSIIS first name	–1.651**	–1.819**	–1.756**	–3.688***
	(0.738)	(0.848)	(0.693)	(0.593)
Jaro-Winkler distance in NYSIIS last name	–0.765***	–0.695***	–0.855***	-0.0696
	(0.253)	(0.243)	(0.272)	(0.144)
Middle initial match, if have one	1.275***	1.661***	1.042***	–
	(0.131)	(0.123)	(0.226)	
Constant	0.109	–0.528	0.0449	0.526
	(0.324)	(0.378)	(0.350)	(0.322)
Observations	21,648	18,690	13,893	29,591

Source: Ward (2018).

Table A.12: Probit coefficients, Part 3

	French	Romanian	Greek	Russian
Year of birth difference = 1	–0.641***	–0.265*	–0.520***	–0.209*
	(0.132)	(0.146)	(0.0841)	(0.110)
Year of birth difference = 2	–1.040***	–0.521***	–0.968***	–0.530***
	(0.158)	(0.153)	(0.103)	(0.118)
Year of birth difference = 3	–0.899***	–0.602***	–1.023***	–0.559***
	(0.158)	(0.161)	(0.115)	(0.124)
Jaro-Winkler distance in first name string	–5.319*	–8.447	–0.214	–2.643
	(2.841)	(5.189)	(1.898)	(3.739)
Jaro-Winkler distance in last name string	–12.31***	–9.571***	–8.833***	–9.095***
	(1.081)	(0.980)	(0.716)	(0.785)
Exact First name match (NYSIIS)	–0.106	–1.240*	0.228	0.366
	(0.358)	(0.712)	(0.287)	(0.558)
Exact first and last name match (NYSIIS)	–1.031***	–0.652***	–0.137	–0.887***
	(0.200)	(0.221)	(0.111)	(0.151)
Total number of hits	–0.349***	–0.240***	–0.209***	–0.223***
	(0.0351)	(0.0306)	(0.0245)	(0.0217)
Total number of hits squared	0.0114***	0.00620***	0.00486***	0.00555***
	(0.00164)	(0.00134)	(0.000802)	(0.000790)
First letter of last name match	0.185	0.103	0.256	0.161
	(0.208)	(0.184)	(0.198)	(0.153)
First letter of first name match	1.283***	0.943*	0.472**	–0.130
	(0.414)	(0.518)	(0.225)	(0.280)
More than two hits have NYSIIS last name match	–0.852***	–0.534	0.895***	0.0920
	(0.323)	(0.362)	(0.140)	(0.258)
One hit has NYSIIS last name match	1.969***	1.759***	0.750***	1.073***
	(0.309)	(0.363)	(0.147)	(0.254)
Jaro-Winkler distance in NYSIIS first name	–3.694***	–2.865***	–3.269***	–5.015***
	(1.024)	(0.846)	(0.670)	(0.820)
Jaro-Winkler distance in NYSIIS last name	0.0116	–0.0180	–0.558***	–0.239
	(0.210)	(0.265)	(0.208)	(0.236)
Middle initial match, if have one	1.179**	-	0.864*	2.620**
	(0.469)		(0.461)	(1.041)
Constant	1.273**	1.849**	0.617	1.320**
	(0.614)	(0.925)	(0.456)	(0.672)
Observations	3,190	2,899	21,761	10,481

Source: Ward (2018).

Table A.13: Probit coefficients, Part 4

	Czech	Polish	Finnish	Hungarian
Year of birth difference = 1	–0.570***	–0.438***	–0.580***	–0.425***
	(0.101)	(0.0742)	(0.0919)	(0.0998)
Year of birth difference = 2	–1.009***	–0.625***	–0.911***	–0.739***
	(0.123)	(0.0861)	(0.109)	(0.111)
Year of birth difference = 3	–1.229***	–0.700***	–0.950***	–1.106***
	(0.148)	(0.0996)	(0.111)	(0.128)
Jaro-Winkler distance in first name string	–6.374**	–1.801	1.632	–7.916***
	(3.125)	(2.694)	(1.783)	(2.525)
Jaro-Winkler distance in last name string	–12.24***	–11.45***	–8.022***	–8.046***
	(0.898)	(0.645)	(0.738)	(0.777)
Exact First name match (NYSIIS)	–0.827*	0.223	1.030***	–0.593*
	(0.432)	(0.364)	(0.294)	(0.356)
Exact first and last name match (NYSIIS)	–0.274*	–0.582***	–0.673***	–0.460***
	(0.162)	(0.112)	(0.122)	(0.132)
Total number of hits	–0.227***	–0.115***	–0.279***	–0.236***
	(0.0265)	(0.0243)	(0.0209)	(0.0225)
Total number of hits squared	0.00560***	0.00154*	0.00784***	0.00572***
	(0.000899)	(0.000793)	(0.000780)	(0.000822)
First letter of last name match	–0.159	0.237*	–0.0230	–0.0458
	(0.157)	(0.140)	(0.135)	(0.154)
First letter of first name match	0.450	0.370*	0.244	0.326
	(0.300)	(0.215)	(0.175)	(0.318)
More than two hits have NYSIIS last name match	0.264	0.300*	0.106	–0.107
	(0.221)	(0.162)	(0.165)	(0.203)
One hit has NYSIIS last name match	1.054***	1.188***	1.396***	1.627***
	(0.222)	(0.165)	(0.159)	(0.202)
Jaro-Winkler distance in NYSIIS first name	–4.807***	–3.273***	–2.033***	–2.655***
	(0.889)	(0.617)	(0.620)	(0.703)
Jaro-Winkler distance in NYSIIS last name	–0.465*	0.0369	–1.251***	–0.139
	(0.276)	(0.236)	(0.467)	(0.191)
Middle initial match, if have one	–0.464	1.537	1.216***	2.947***
	(2.109)	(4.076)	(0.350)	(1.072)
Constant	2.914***	0.863*	0.304	1.625***
	(0.615)	(0.482)	(0.382)	(0.490)
Observations	16,041	27,298	8,006	9,891

Source: Ward (2018).

Table A.14: Critical values used to keep links

Language	Probability Threshold	Ratio of First-Best Score to Second-Best Score	PPV PPV	TPR TPR
English	0.305	1.4	0.904	0.728
German	0.434	1.2	0.901	0.714
Yiddish, Jewish	0.372	1.7	0.901	0.594
Dutch	0.337	1.1	0.901	0.881
Swedish	0.268	3.4	0.901	0.572
Danish	0.356	1.9	0.901	0.611
Norwegian	0.331	1.5	0.902	0.731
Italian	0.521	1.5	0.901	0.432
French	0.313	1.2	0.903	0.871
Romanian	0.402	1.6	0.905	0.643
Greek	0.527	2.2	0.904	0.285
Russian	0.397	4.3	0.904	0.479
Czech/Slovak	0.325	3.1	0.901	0.622
Polish	0.357	9.1	0.904	0.383
Finnish	0.257	2.4	0.900	0.688
Hungarian	0.38	7.7	0.903	0.518

Notes: This table gives the meta-parameters for inclusion in the linked sample. The predicted probability for a match must be above the probability threshold, and the predicted probability must be at least the multiple (in Column 3) of the second-best score. The PPV, or positive prediction value, is the ratio of true positives to all positives; a higher number indicates fewer false positives. The TPR, or the true positive rate, is the ratio of true positives to all possible links; a lower number reflects that the probit does not include all matches from the hand linked data.
Source: Ward (2018)

Table A.15: Robustness of Table 2.3: Effect of age at arrival on occupations

Age at Arrival:	White-Col.	Skilled	Farmer	Unskilled	Log (Occ. Score)	
					1940 Census	1950 Occscore
2 to 3	–0.00301	–0.00558	–0.00739	0.0160	–0.0477***	–0.0137
	(0.0251)	(0.0235)	(0.0103)	(0.0275)	(0.0174)	(0.0195)
4 to 5	–0.00722	–0.00405	–0.00397	0.0152	–0.0528***	–0.0160
	(0.0255)	(0.0236)	(0.0106)	(0.0280)	(0.0184)	(0.0203)
6 to 7	–0.0175	–0.00943	–0.0113	0.0382	–0.0930***	–0.0294
	(0.0269)	(0.0250)	(0.0112)	(0.0300)	(0.0185)	(0.0210)
8 to 9	–0.0260	0.0161	–0.0158	0.0258	–0.111***	–0.0254
	(0.0271)	(0.0250)	(0.0112)	(0.0297)	(0.0188)	(0.0213)
10 to 11	–0.0362	0.00340	–0.0143	0.0471	–0.118***	–0.0521**
	(0.0280)	(0.0265)	(0.0123)	(0.0314)	(0.0201)	(0.0227)
12 to 13	–0.0237	0.0111	–0.0203	0.0329	–0.140***	–0.0539**
	(0.0323)	(0.0306)	(0.0134)	(0.0362)	(0.0226)	(0.0258)
14 to 15	–0.0722**	0.0317	–0.0211	0.0616*	–0.151***	–0.0730***
	(0.0323)	(0.0318)	(0.0141)	(0.0369)	(0.0224)	(0.0256)
16 to 17	–0.0773**	0.00595	–0.0185	0.0899**	–0.156***	–0.0834***
	(0.0348)	(0.0339)	(0.0163)	(0.0389)	(0.0242)	(0.0290)
18 to 20	–0.0703*	0.0328	–0.0303	0.0678	–0.148***	–0.0610*
	(0.0394)	(0.0379)	(0.0195)	(0.0441)	(0.0276)	(0.0316)
N	20,715	20,715	20,715	20,715	20,715	20,715
R^2	0.591	0.554	0.677	0.584	0.712	0.624

Notes and Source: This table recreates Table 2.3 from the main text, but with the sample linked using the Feigenbaum (2016) method.

Table A.16: Robustness of Table 2.4: The return to education and experience

	Full Sample	Only NW Europe	Only SE Europe	Full Sample
US Educ.	0.0611***	0.0705***	0.0501***	0.0705***
	(0.0102)	(0.0117)	(0.0181)	(0.0120)
US Educ. ×				–0.0203
SE Europe				(0.0211)
Foreign Educ.	0.0468***	0.0614***	0.0354**	0.0614***
	(0.00856)	(0.0105)	(0.0141)	(0.0107)
Foreign Educ. ×				–0.0259
SE Europe				(0.0173)
Foreign Exp.	0.0289	0.0438	0.0290	0.0438
	(0.0235)	(0.0490)	(0.0318)	(0.0501)
Foreign Exp. ×				–0.0148
SE Europe				(0.0587)
(Foreign Exp./10)^2	–0.170	–0.607	–0.206	–0.607
	(0.274)	(0.820)	(0.329)	(0.839)
(Foreign Exp./10)^2				0.402
× SE Europe				(0.897)
US Exp.	0.0792***	0.0943***	0.0574**	0.0943***
	(0.0123)	(0.0138)	(0.0232)	(0.0141)
US Exp. ×				–0.0369
SE Europe				(0.0264)
(US Exp./10)^2	–0.128***	–0.172***	–0.0751*	–0.172***
	(0.0232)	(0.0259)	(0.0419)	(0.0265)
(US Exp./10)^2 ×				0.0973**
SE Europe				(0.0482)
Observations	14,703	8,980	5,723	14,703
R^2	0.715	0.705	0.720	0.717

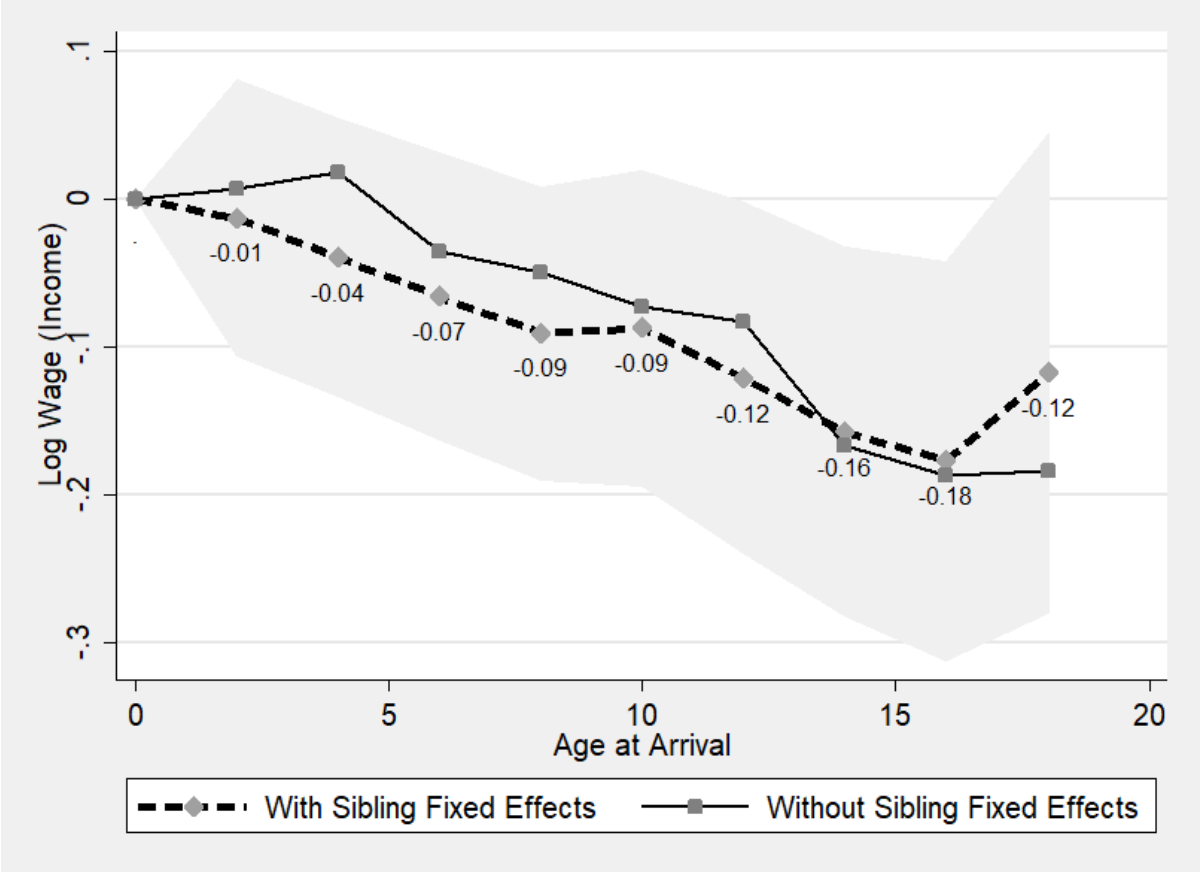
Notes and Source: This table recreates Table 2.4 from the main text, but with the sample linked using the Feigenbaum (2016) method.

Table A.17: Robustness of Table 2.5: Effect of age at arrival on social outcomes

	Intermarriage		Spatial Assimilation	
Age at arrival:	Native spouse	Spouse from Different source	Fraction of county native HH	Fraction of census page native HH
2 to 3	–0.0176	–0.0357	0.00321	0.000183
	(0.0483)	(0.0430)	(0.00520)	(0.0110)
4 to 5	–0.0649	–0.0583	–0.00333	–0.00929
	(0.0502)	(0.0447)	(0.00552)	(0.0113)
6 to 7	–0.0744	–0.0832*	–0.00155	0.00463
	(0.0518)	(0.0470)	(0.00584)	(0.0123)
8 to 9	–0.142***	–0.156***	0.00100	0.00411
	(0.0512)	(0.0458)	(0.00576)	(0.0119)
10 to 11	–0.178***	–0.192***	0.00149	0.0107
	(0.0539)	(0.0481)	(0.00617)	(0.0127)
12 to 13	–0.244***	–0.249***	0.00303	0.0172
	(0.0620)	(0.0571)	(0.00703)	(0.0152)
14 to 15	–0.307***	–0.308***	0.00400	–0.00113
	(0.0619)	(0.0567)	(0.00709)	(0.0151)
16 to 17	–0.382***	–0.381***	0.00867	–2.41e-05
	(0.0670)	(0.0621)	(0.00781)	(0.0164)
18 to 19	–0.424***	–0.443***	0.00893	–0.0106
	(0.0735)	(0.0707)	(0.00934)	(0.0187)
Observations	12,503	12,503	21,994	21,994
R^2	0.732	0.741	0.737	0.658

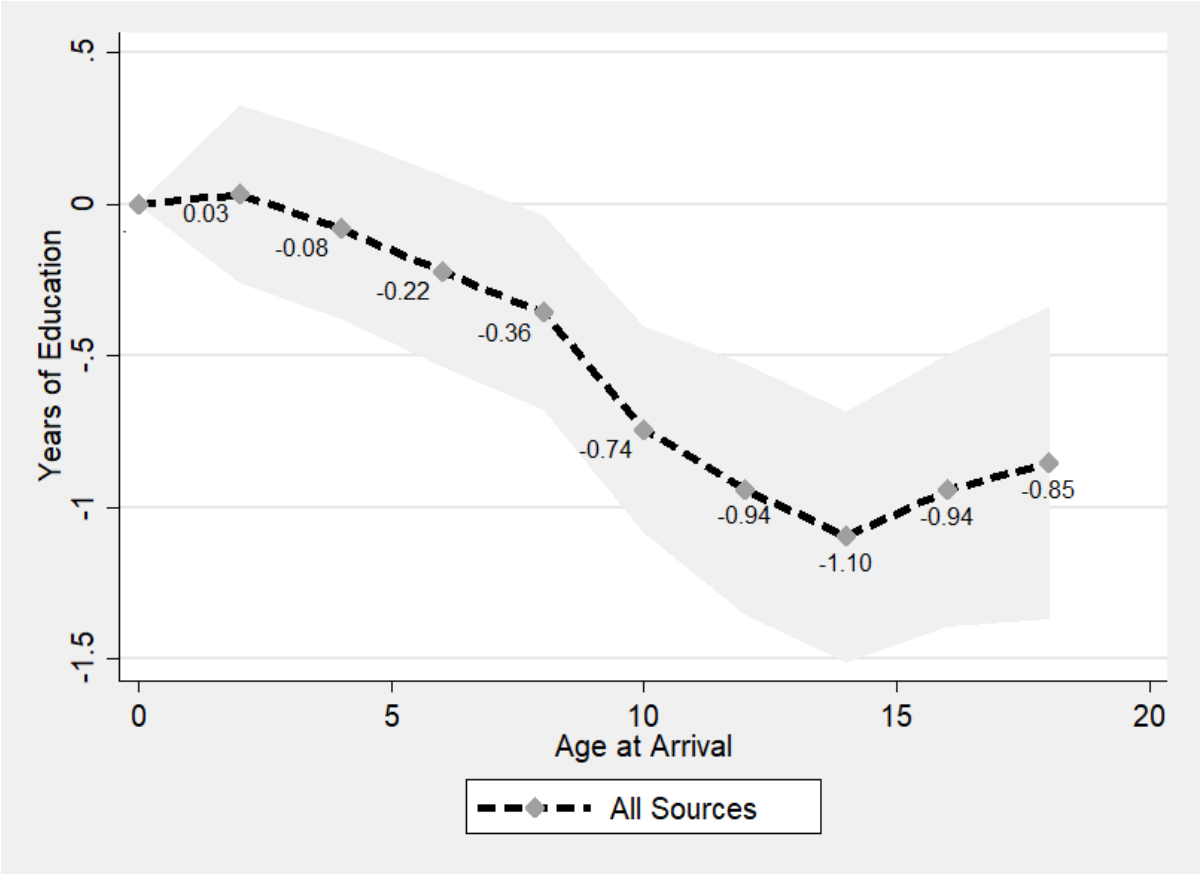
Notes and Source: This table recreates Table 2.5 from the main text, but with the sample linked using the Feigenbaum (2016) method.

Figure A.2: Robustness of Figure 2.3: The negative effect of age at arrival on the native-immigrant gap in wage income in 1940



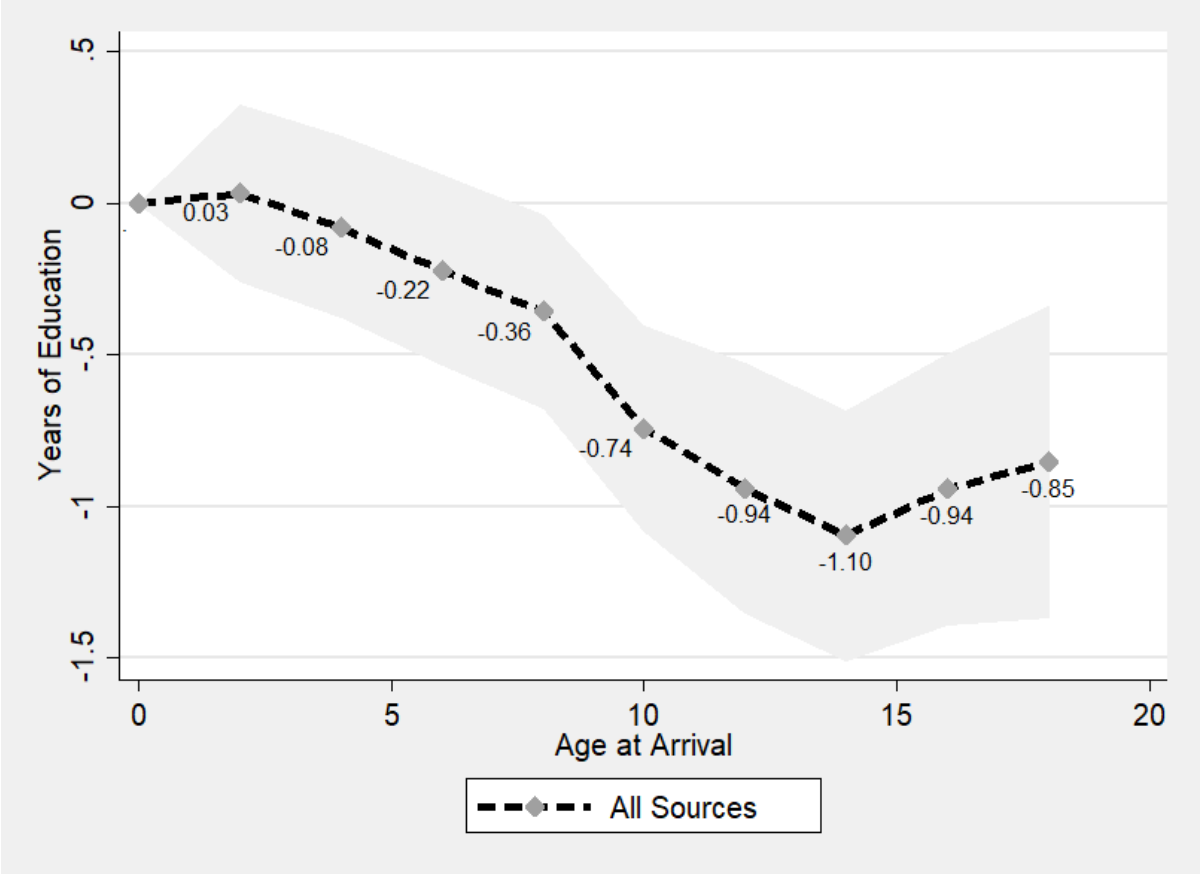
Notes: The dependent variable is the age-adjusted gap in log wage income between immigrants and natives. Self-employed workers are dropped. The figure shows the estimated fixed effects for age at arrival with age at arrival of zero and one being the excluded group. The shaded area is the 95 percent confidence interval when using sibling fixed effects. Standard errors are clustered at the household level.
Sources: Sample of brothers linked from Ellis Island records to the 1940 Census using Feigenbaum (2016) method.

Figure A.3: Robustness of Figure 2.4: The negative effect of age at arrival on the native-immigrant gap in years of education in 1940



Notes: The dependent variable is the age-adjusted gap in years of education between immigrants and natives. The figure shows the estimated fixed effects for age at arrival with age at arrival of zero and one being the excluded group. The shaded area is the 95 percent confidence interval. Standard errors are clustered at the household level.
Sources: Linked sample of brothers from Ellis Island records to the 1940 Census using Feigenbaum (2016) method.

Figure A.4: Robustness of Figure 2.5: The age-at-arrival profiles were differently sloped across new and old sources



Notes: The figure shows the estimated fixed effects for age at arrival with age at arrival of zero and one being the excluded group. The shaded area is the 95 percent confidence interval for the New Source group. Standard errors are clustered at the household level. New source countries are in Southern and Western Europe and Old Source countries are in Northern and Western Europe.
Sources: Linked sample of brothers from Ellis Island records to the 1940 Census using Feigenbaum (2016)method.

Bibliography

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. (2017). When Should You Adjust Standard Errors for Clustering? Working Paper 24003, NBER Working Paper.

Abramitzky, R. and Boustan, L. (2017). Immigration in American Economic History. *Journal of Economic Literature*, 55(4):1311–1345.

Abramitzky, R., Boustan, L., and Eriksson, K. (2012). Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration. *American Economic Review*, 102(5):1832–1856.

Abramitzky, R., Boustan, L., and Eriksson, K. (2013). Have the Poor Always been Less Likely to Migrate? Evidence from Inheritance Practices during the Age of Mass Migration. *Journal of Development Economics*, 102:2–14.

Abramitzky, R., Boustan, L., and Eriksson, K. (2014). A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration. *Journal of Political Economy*, 122(3):467–506.

Abramitzky, R., Boustan, L., and Eriksson, K. (2016). Cultural Assimilation during the Age of Mass Migration. Working Paper 22381, NBER Working Paper.

ABS (1996). Feature Article – Early Tasmanian Settlements. In *1301.6 – Tasmanian Year Book, 1996*. Australian Bureau of Statistics, Canberra.

ABS (2002). Feature Article – History of Tasmania's Population 1803–2000. In *1384.6 – Statistics – Tasmania*. Australian Bureau of Statistics, Canberra.

Alesina, A. and Spolaore, E. (2003). *The Size of Nations*. MIT Press, Cambridge.

Alesina, A., Spolaore, E., and Wacziarg, R. (2000). Economic Integration and Political Disintegration. *American Economic Review*, 90(5):1276–1296.

Alexander, R. and Ward, Z. (2018). Age at Arrival and Assimilation during the Age of Mass Migration. Technical report, Inter-university Consortium for Political and Social Research [distributor].

Alien Immigration (1898). Alien Immigration. *Clarence River Advocate (21 June 1898)*.

Almond, D., Currie, J., and Duque, V. (2017). Childhood Circumstances and Adult Outcomes: Act II. Working Paper 23017, NBER Working Paper.

Åslund, O., Böhlmark, A., and Skans, O. N. (2015). Childhood and Family Experiences and the Social Integration of Young Migrants. *Labour Economics*, 35:135–144.

Atkinson, A. (2013). Federation, Democracy and the Struggle against a Single Australia. *Australian Historical Studies*, 44:169–188.

Bailey, M., Cole, C., Henderson, M., and Massey, C. (2017). How Well Do Automated Methods Perform in Historical Samples? Evidence from New Ground Truth. Working Paper 24019, NBER Working Paper.

Baines, D. (1995). *Emigration from Europe 1815–1930*. Cambridge University Press, New York.

Baker, M. and Benjamin, D. (1994). The Performance of Immigrants in the Canadian Labor Market. *Journal of Labor Economics*, 12:369–405.

Bandiera, O., Mohnen, M., Rasul, I., and Viarengo, M. (2016). Nation-Building Through Compulsory Schooling During the Age of Mass Migration. Sticerd—Economic Organisation and Public Policy Discussion Papers Series, Suntory and Toyota International Centres for Economics and Related Disciplines, LSE.

Bandiera, O., Rasul, I., and Viarengo, M. (2013). The Making of Modern America: Migratory Flows in the Age of Mass Migration. *Journal of Development Economics*, 102:23–47.

Bannon, J. (1999). South Australia. In Irving, H., editor, *The Centenary Companion to Australian Federation*. Cambridge University Press, Cambridge.

Barnard, A. (1985). Colonial and State Government Finances. Source Papers 8, 9, 10, 14, 15, 16, Centre for Economic History.

Barone, G. and Mocetti, S. (2016). Intergenerational Mobility in the Very Long Run: Florence 1427–2011. Technical report, Banca D'Italia (Bank of Italy), Working papers, Number 1060, April 2016.

Bastin, J. (1951). Federation and Western Australia: A Contribution to the Parker-Blainey Discussion. *Historical Studies: Australia and New Zealand*, 5:47–58.

Biavaschi, C., Giulietti, C., and Siddique, Z. (2017). The Economic Payoff of Name Americanization. *Journal of Labor Economics*, 35(4):1089–1116.

Birrell, R. (1995). *A Nation of Our Own: Citizenship and Nation-Building in Federation Australia*. Longman, Melbourne.

Blackton, C. S. (1958). Australian Nationality and Nativism: The Australian Natives' Association, 1885–1901. *Journal of Modern History*, 30:37–46.

Blainey, G. (1950). The Role of Economic Interests in Australian Federation: A Reply to Professor R. S. Parker. *Historical Studies: Australia and New Zealand*, 4:224–237.

Bleakley, H. and Chin, A. (2004). Language Skills and Earnings: Evidence from Childhood Immigrants. *Review of Economics and Statistics*, 86(2):481–496.

Bleakley, H. and Chin, A. (2010). Age at Arrival, English Proficiency, and Social Assimilation among U.S. Immigrants. *American Economic Journal. Applied Economics*, 2(1):165–192.

Böhlmark, A. (2008). Age at Immigration and School Performance: A Siblings Analysis Using Swedish Register Data. *Labour Economics*, 15(6):1366–1387.

Bolton, G. (1963). *A Thousand Miles Away: A History of North Queensland to 1920*. Jacaranda Press, Canberra.

Bolton, P. and Roland, G. (1997). The Breakup of Nations: A Political Economy Analysis. *Quarterly Journal of Economics*, 112(4):1057–1090.

Borjas, G. J. (1985). Assimilation, Changes in Cohort Quality, and the Earnings of Immigrants. *Journal of Labor Economics*, 3(4):463–489.

Borjas, G. J. (1987). Self-Selection and the Earnings of Immigrants. *The American Economic Review*, 77(4):531–553.

Borjas, G. J. (1994). Long-Run Convergence of Ethnic Skill Differentials: The Children and Grandchildren of the Great Migration. *ILR Review*, 47(4):553–573.

Bradley, J., Kippen, R., Maxwell-Stewart, H., McCalman, J., and Silcot, S. (2010). Research Note: The Founders and Survivors Project. *History of the Family*, 15:467–477.

Brown, A. (2004). One Continent, Two Federalisms: Rediscovering the Original Meanings of Australian Federal Ideas. *Australian Journal of Political Science*, 39(3):485–504.

Butlin, N. (1985). Australian National Accounts 1788–1983. Technical report, Source Papers in Economic History Number 6, Australian National University, Canberra.

Cahill, A. E. (2001). Catholics and Australian Federation. *Journal of the Australian Catholic Historical Society*, 22:9–30.

Camm, J., McQuilton, J., and Yorke, S. (1983). *Australian Census Maps 1901 [Cartographic Material]. Historical Geography Monograph No. 1 Australia 1788–1988*. A Bicentennial History.

Card, D., DiNardo, J., and Estes, E. (2000). The More Things Change: Immigrants and the Children of Immigrants in the 1940s, the 1970s, and the 1990s. In *Issues in the Economics of Immigration*, pages 227–270. University of Chicago Press.

Carneiro, P. M., Lee, S., and Reis, H. (2015). Please call me John: name choice and the assimilation of immigrants in the United States, 1900-1930. CeMMAP working papers CWP28-15, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Chetty, R. and Hendren, N. (2017a). The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects. Working Paper 23001, NBER Working Paper.

Chetty, R. and Hendren, N. (2017b). The Impacts of Neighborhoods on Intergenerational Mobility II: County Level Estimates. Working Paper 23002, NBER Working Paper.

Chetty, R., Hendren, N., and Katz, L. F. (2016). The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment. *American Economic Review*, 106(4):855–902.

Chiswick, B. R. (1978). The Effect of Americanization on the Earnings of Foreign-Born Men. *Journal of Political Economy*, 86(5):897–921.

Clark, G. (2014). *The Son Also Rises: Surnames and the History of Social Mobility*. Princeton University Press.

Clark, G. and Cummins, N. (2013). Surnames and Social Mobility: England 1230–2012. Technical report, Department of Economic History, The London School of Economics.

Clark, G., Leigh, A., and Pottenger, M. (2017). Immobile Australia: Surnames Show Strong Status Persistence, 1870-2017. Technical report, CESifo Working Paper Series.

Clarke, A. (2016). Age at Immigration and the Educational Attainment of Foreign-Born Children in the United States: The Confounding Effects of Parental Education. *International Migration Review*.

Clay, K., Lingwall, J., and Jr, M. S. (2016). Laws, Educational Outcomes, and Returns to Schooling: Evidence from the Full Count 1940 Census. Working Paper 22855, NBER Working Paper.

Cohn, R. L. (2009). *Mass Migration under Sail: European Immigration to the Antebellum United States*. Cambridge University Press, Cambridge.

Coleman, W. (2017). The Social and Economic Determinants of Voting 'Yes' in South Australia's Federation Referenda. Working Paper DP700, Centre for Economic Policy Research.

Collins, W. J. and Wanamaker, M. H. (2014). Selection and Economic Gains in the Great Migration of African Americans: New Evidence from Linked Census Data. *American Economic Journal: Applied Economics*, 6(1):220–252.

Collins, W. J. and Wanamaker, M. H. (2017). Up from Slavery? African American Intergenerational Economic Mobility Since 1880. Working Paper NBER, NBER Working Paper.

Cunha, F., Heckman, J. J., Lochner, L., and et al (2006). Interpreting the Evidence on Life Cycle Skill Formation. *Handbook of the Economics of Education*, 1:697–812.

Cutler, D. M., Glaeser, E. L., and Vigdor, J. L. (2008). Is the Melting Pot Still Hot? Explaining the Resurgence of Immigrant Segregation. *Review of Economics and Statistics*, 90(3):478–497.

de Garis, B. (1999). The Colonies' Paths to Federation: Western Australia. In Irving, H., editor, *The Centenary Companion to Australian Federation*. Cambridge University Press, Cambridge.

de Looper, M. (2014). *Death Registration and Mortality Trends in Australia 1856-1906*. PhD thesis, The Australian National University, Canberra.

Deakin, A. (1944). *The Federal Story: The Inner History of the Federal Cause*. Robertson & Mullens.

Desmet, K., Le Breton, M., Ortuño-Ortín, I., and Weber, S. (2011). The Stability and Breakup of Nations: A Quantitative Analysis. *Journal of Economic Growth*, 16(3):183–213.

Deutscher, N. and Mazumder, B. (2019). Intergenerational Mobility in Australia: National

and Regional Estimates Using Administrative Data. Technical report, Life Course Australia.

Eddy, J. J. (1978a). Imperial Sentiment as a Factor in Centralising Australian Federation. In Hodgins, B. W., Wright, D., and Heick, W. H., editors, *Federalism in Canada and Australia: The Early Years*. Australian National University, Canberra.

Eddy, J. J. (1978b). Politics in New South Wales: The Federation Issue and the Move Away from Faction and Parochialism. In Hodgins, B. W., Wright, D., and Heick, W. H., editors, *Federalism in Canada and Australia: The Early Years,*. Australian National University, Canberra.

Emigration from the United Kingdom (1838). Emigration from the United Kingdom. *Journal of the Statistical Society of London,*, 1(3):155–167.

Eriksson, K. and Niemesh, G. (2016). Death in the Promised Land: The Great Migration and Black Infant Mortality. Technical Report 3071053, SSRN.

Federation and Taxation (1899). Federation and Taxation. *North Western Advocate and the Emu Bay Times (28 April 1899)*.

Feigenbaum, J. J. (2016). Automated Census Record Linking: A Machine Learning Approach. Working Paper.

Forster, C. (1977). Federation and the Tariff. *Australian Economic History Review*, 17:95–116.

Fowler, A. (2013). Electoral and Policy Consequences of Voter Turnout: Evidence from Compulsory Voting in Australia. *Quarterly Journal of Political Science*, 8(2):159–182.

Friedberg, R. M. (1992). The Labor Market Assimilation of Immigrants in the United States: The Role of Age at Arrival. *Brown University*.

Friedberg, R. M. (2000). You Can't Take It With You? Immigrant Assimilation and the Portability of Human Capital. *Journal of Labor Economics*, 18(2):221–251.

Goldin, C. and Margo, R. A. (1992). The Great Compression: The Wage Structure in the United States at Mid-century. *The Quarterly Journal of Economics*, 107(1):1–34.

Gould, J. D. (1980). European International Emigration: The Role of Diffusion and Feedback. *Journal of European Economic History*, 9:267–315.

Greenwood, M. J. (2007). Modeling the Age and Age Composition of Late Nineteenth Century U.S. Immigrants from Europe. *Explorations in Economic History*, 44(2):255–269.

Güell, M., Rodríguez Mora, J. V., and Telmer, C. I. (2014). The Informational Content of Surnames, the Evolution of Intergenerational Mobility, and Assortative Mating. *The Review of Economic Studies*, 82(2):693–735.

Hatton, T. J. (1997). The Immigrant Assimilation Puzzle in Late Nineteenth-Century America. *The Journal of Economic History*, 57(1):34–62.

Hatton, T. J. and Williamson, J. G. (1998). *The Age of Mass Migration: Causes and Economic Impact*. Oxford University Press, Oxford, UK.

Hewett, P. (1969). Aspects of Campaigns in South Eastern New South Wales at the Federation Referenda of 1898 and 1899. In Martin, A. W., editor, *Essays in Australian Federation,*. Melbourne University Press, Melbourne.

Higham, J. (1955). *Strangers in the Land: Patterns of American Nativism, 1860–1925*. Rutgers University Press, Brunswick, NJ.

Hillman, W. (1978). The 1900 Federal Referendum in Western Australia. *Studies in Western Australia History*, 2:51–65.

Huntington, S. P. (2004). *Who Are We?: The Challenges to America's National Identity*. Simon and Schuster, New York.

Hutchinson, E. P. (1958). Notes on Immigration Statistics of the United States. *Journal of the American Statistical Association*, 53(284):963–1025.

Irving, H. (1999). *To Constitute a Nation: A Cultural History of Australia's Constitution*. Cambridge University Press, Cambridge.

Irwin, D. A. (2006). The Impact of Federation on Australia's Trade Flows. *Economic Record*, 82:315–324.

Kippen, R. (2002a). An Indispensable Duty of Government: Civil Registration in Nineteenth-Century Tasmania. *Tasmanian Historical Studies*, 8(1):42–58.

Kippen, R. (2002b). *Death in Tasmania: Using Civil Registers to Measure Nineteenth-Century Cause-Specific Mortality*. PhD thesis, The Australian National University, Canberra.

Korobacz, V. (1971). The Legislative Council of Van Diemen's Land 1825–1856. Master's thesis, University of Tasmania,.

Leigh, A. (2007). Intergenerational Mobility in Australia. *The B.E. Journal of Economic Analysis & Policy*, 7(2):1–26.

Lindahl, M., Palme, M., Massih, S. S., and Sjögren, A. (2015). Long-Term Intergenerational Persistence of Human Capital: An Empirical Analysis Of Four Generations. *Journal of Human Resources*, 50(1):1–33.

Lindert, P. H. (2004). *Growing Public: Social Spending and Economic Growth Since the Eighteenth Century, Volume 1: The Story*. Cambridge University Press, Cambridge, UK.

Lloyd, P. (2015). Customs Union and Fiscal Union in Australia at Federation. *Economic Record*, 91:155–171.

Lloyd, P. (2017). Excise Tax Harmonisation in Australia at Federation. *Australian Economic History Review*, 57:45–64.

Logan, T. D. and Parman, J. M. (2017). The National Rise in Residential Segregation. *Journal of Economic History*, 77(1):127–170.

Loveday, P. (1972). The Federal Convention, an Analysis of the Voting. *Australian Journal of Politics and History*, 18:169–188.

Martin, G. (2001). *Australia, New Zealand and Federation 1883–1901*. Menzies Centre for Australian Studies.

Massey, C. G. (2017). Playing with Matches: An Assessment of Accuracy in Linked Historical Data. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 50(3):129–143.

McAllister, I. (2001). Elections Without Cues: The 1999 Australian Republic Referendum. *Australian Journal of Political Science*, 36(2):247–269.

McLean, I. W. (2013). *Why Australia Prospered: The Shifting Sources of Economic Growth*. Princeton University Press.

Meng, X. and Gregory, R. G. (2005). Intermarriage and the Economic Assimilation of Immigrants. *Journal of Labor Economics*, 23(1):135–174.

Meredith, D. and Oxley, D. (2014). The Convict Economy. In Ville, S. and Withers, G., editors, *The Cambridge Economic History of Australia*, pages 97–122. Cambridge University Press.

Merrett, D. T. (2013). The Australian Bank Crashes of the 1890s Revisited. *Business History Review*, 87(3):407–429.

Minns, C. (2000). Income, Cohort Effects, and Occupational Mobility: A New Look At Immigration to the United States at the Turn of the Twentieth Century. *Explorations in Economic History*, 37(4):326–350.

Montalvo, J. G. and Reynal-Querol, M. (2005). Ethnic Polarization, Potential Conflict, and Civil Wars. *American Economic Review*, 95(3):796–816.

Moyle, H. (2015). *The Fall of Fertility in Tasmania in the Late 19th and Early 20th Centuries*. PhD thesis, The Australian National University, Canberra.

Norris, R. (1969). Economic Influences on the 1898 South Australian Referendum. In Martin, A. W., editor, *Essays in Australian Federation*. Melbourne University Press, Melbourne.

Norris, R. (1978). Towards a Federal Union. In Hodgins, B. W., Wright, D., and Heick, W. H., editors, *Federalism in Canada and Australia: The Early Years*. Australian National University, Canberra.

Olivetti, C. and Paserman, M. D. (2015). In the Name of the Son (and the Daughter): Intergenerational Mobility in the United States, 1850–1940. *American Economic Review*, 105(8):2695–2724.

Olivetti, C., Paserman, M. D., and Salisbury, L. (2018). Three-Generation Mobility in the United States, 1850–1940: The Role Of Maternal And Paternal Grandparents. *Explorations in Economic History*, 70:73–90.

Parker, R. S. (1949). Australian Federation: The Influence of Economic Interests and Political Pressures. *Historical Studies: Australia and New Zealand*, 4:1–24.

Parker, R. S. (1950). Some Comments on the Role of Economic Interests in Australian Federation. *Historical Studies: Australia and New Zealand*, 4:238–240.

Pettman, J. (1969). The Australian Natives Association and Federation in South Australia. In Martin, A. W., editor, *Essays in Australian Federation*. Melbourne University Press, Melbourne.

Preston, S. H. and Haines, M. R. (1991). *Fatal Years: Child Mortality in Late Nineteenth-Century America*. Princeton University Press.

Pringle, R. (1978). Public Opinion in the Federal Referendum Campaigns in New South Wales, 1898–1899. *Journal of the Royal Australian Historical Society*, 64:235–251.

Quick, J. and Garran, R. R. (1901). *The Annotated Constitution of the Australian Commonwealth*. Angus & Robertson.

Reynolds, H. (1969). Men of Substance and Deservedly Good Repute: The Tasmanian Gentry 1856–1875. *Australian Journal of Politics & History*, 15(3):61–72.

Rhodes, G. (1988). *The Australian Federation Referenda 1898–1900 A Spatial Analysis of Voting Behaviour*. PhD thesis, London School of Economics.

Rhodes, G. (2002). *Votes for Australia: How Colonials Voted at the 1898–1900 Federation Referendums*. Griffith University, Centre for Australian Public Sector Management, Brisbane.

Ruggles, S., Genadek, K., Goeken, R., Grover, J., and Sobek, M. (2017). Minnesota Population Center. Integrated Public Use Microdata Series, International: Version 7.1 [dataset]. Technical report, Minneapolis: University of Minnesota.

Sawer, M. (2001). Women and Government in Australia. In *Year Book Australia, 2001, Cat. No. 1301.0*. ABS, Canberra.

Schaafsma, J. and Sweetman, A. (2001). Immigrant Earnings: Age at Immigration Matters. *Canadian Journal of Economics*, 34(4):1066–99.

Schoellman, T. (2016). Early Childhood Human Capital and Development. *American Economic Journal: Macroeconomics*, 8(3):145–74.

School Fees (1846). School Fees. *Hobart Courier (4 July 1846)*.

Solon, G. (1992). Intergenerational Income Mobility in the United States. *American Economic Review*, 82(3):393–408.

Solon, G. (2015). What Do We Know So Far about Multigenerational Mobility? In *paper prepared for HCEO Conference on Social Mobility*. available at solon/MultigenJan2015.pdf.

Spitzer, Y. and Zimran, A. (2017). Migrant Self-Selection: Anthropometric Evidence from the Mass Migration of Italians to the United States, 1907–1925. *Mimeo*.

Spolaore, E. (2016). The Economics of Political Borders. In Kontorovich, E. and Parisi, F., editors, *Economic Analysis of International Law*, pages 11–43. Edward Elgar.

US Immigration Commission (1910). The Children of Immigrants in Schools, Vol. 1. Reports of the Immigration Commission, 61st Congress, 3rd Session, Washington, DC.

Van den Berg, G. J., Lundborg, P., Nystedt, P., and Rooth, D.-O. (2014). Critical Periods During Childhood and Adolescence. *Journal of the European Economic Association*, 12(6):1521–1557.

Walker, F. A. (1896). Restriction of Immigration. *Atlantic Monthly*, 77:822–829.

Ward, Z. (2017). Birds of Passage: Return Migration, Self-Selection, and Immigration Quotas. *Explorations in Economic History*, 64:37–52.

Ward, Z. (2018). Have Language Skills Always Been So Valuable? The Low Return to English Fluency During the Age of Mass Migration. Working Paper.

Warden, D. (1999). Tasmania. In Irving, H., editor, *The Centenary Companion to Australia Federation*. Cambridge University Press, Cambridge.

Williamson, J. G. (1995). The Evolution of Global Labor Markets Since 1830: Background Evidence and Hypotheses. *Explorations in Economic History*, 32:141–196.

Wise, B. R. (1913). *The Making of the Australian Commonwealth*. Longmans, Green.

Womanhood Suffrage (1897). Womanhood Suffrage. *Dawn (1 Apr 1897)*, page 14.